SGD at Small Learning Rates

Samuel C. Tenka COLI@MIT.EDU
MIT, CSAIL

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We quantify how gradient noise shapes the dynamics of stochastic gradient descent (SGD) by taking Taylor series in the learning rate. We present in particular a new diagram-based notation that permits resummation to convergent results. We employ our theory to contrast SGD against two popular approximations: deterministic descent and stochastic differential equations. We find that SGD's trajectory avoids both gradient noise and minima that are sharp with respect to gradient noise. Paired with our results that connect overfitting to the width of minima, these repulsions suggest a mechanism for the unexpected generalization of overparameterized learning scenarios.

Keywords: SGD, learning rates, generalization, gradient noise.

1. Introduction

1.1. Questions about SGD

1.2. Background, notation, assumptions

1.3. Related work

It was Kiefer and Wolfowitz (1952) who, in uniting gradient descent (Cauchy, 1847) with stochastic approximation (Robbins and Monro, 1951), invented SGD. Since the development of backpropagation for efficient differentiation (Werbos, 1974), SGD has been used to train connectionist models, e.g. neural networks (Bottou, 1991), recently to remarkable success (LeCun et al., 2015).

Several lines of work treat the overfitting of SGD-trained networks (Neyshabur et al., 2017a). For example, Bartlett et al. (2017) controls the Rademacher complexity of deep hypothesis classes, leading to optimizer-agnostic generalization bounds. Yet SGD-trained networks generalize despite their ability to shatter large sets (Zhang et al., 2017), so generalization must arise from not only architecture but also optimization (Neyshabur et al., 2017b). Others approximate SGD by SDE to analyze implicit regularization (e.g. Chaudhari and Soatto (2018)), but, per Yaida (2019), such continuous-time analyses cannot treat SGD noise correctly. We avoid these pitfalls by Taylor expanding around $\eta = 0$ as in Roberts (2018). Unlike that work, we generalize beyond order η^1 and $T = 2$,. To support this generalization, we develop new summation techniques with improved large-T convergence. Our interpretion of the resulting terms offers a new qualitative picture of SGD as a superposition of several simpler information-flow processes.

Our predictions are vacuous for large η. Other analyses treat large-η learning phenomenologically, whether by finding empirical correlates of gen. gap (Liao et al., 2018), by showing that *flat* minima generalize (Hoffer et al. (2017), Keskar et al. (2017), Wang et al. (2018)), or by showing that *sharp* minima generalize (Stein (1956), Dinh et al. (2017), Wu et al. (2018)). Our theory reveals that SGD's implicit regularization mediates between these seemingly clashing intuitions.

Prior work analyzes SGD perturbatively: Dyer and Gur-Ari (2019) perturb in inverse network width, using 't Hooft diagrams to correct the Gaussian Process approximation for specific deep nets. Perturbing to order η^2, Chaudhari and Soatto (2018) and Li et al. (2017) are forced to assume uncorrelated Gaussian noise. By contrast, we use Penrose diagrams to compute test losses to arbitrary order in η. We allow correlated, non-Gaussian noise and thus *any* smooth architecture. For instance, we do not

2. Perturbative theory of SGD

2.1. Trivial example

2.2. Perturbation as technique

2.3. Insights from diagrams

2.4. Resummation

3. Consequences of the theory

3.1. SGD descends on a C-smoothed landscape and prefers minima flat w.r.t. C.

3.2. Both flat and sharp minima overfit less

3.3. High-C regions repel small-(E, B) SGD more than large-(E, B) SGD

3.4. Non-Gaussian noise affects SGD but not SDE

4. Experiments

4.1. A

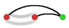
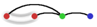
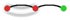
4.2. B

4.3. C

4.4. Overfitting and the width of minima

5. Conclusion

We presented a diagram-based method for studying stochastic optimization on short timescales or near minima. Corollaries **??** and **??** together offer insight into SGD's success in training deep networks: SGD avoids curvature and noise, and curvature and noise control generalization.

Analyzing , we proved that **flat and sharp minima both overfit less** than medium minima. Intuitively, flat minima are robust to vector noise, sharp minima are robust to covector noise, and medium minima robust to neither. We thus proposed a regularizer enabling gradient-based hyperparameter tuning. Inspecting , we extended Wei and Schwab (2019) to nonconstant, nonisotropic covariance to reveal that **SGD descends on a landscape smoothed by the current covariance** C. As C evolves, the smoothed landscape evolves, resulting in non-conservative dynamics. Examining , we showed that **GD may emulate SGD**, as conjectured by Roberts (2018). This is significant because, while small batch sizes can lead to better generalization (Bottou, 1991), modern infrastructure increasingly rewards large batch sizes (Goyal et al., 2018).

Since our predictions depend only on loss data near initialization, they break down after the weight moves far from initialization. Our theory thus best applies to small-movement contexts, whether for long times (large ηT) near an isolated minimum or for short times (small ηT) in general. Thus, the theory might help to analyze meta-learners based on fine-tuning (e.g. Finn et al. (2017)'s MAML).

Much as meteorologists understand how warm and cold fronts interact despite long-term forecasting's intractability, we quantify how curvature and noise contribute to counter-intuitive dynamics governing each short-term interval of SGD's trajectory. Equipped with our theory, practitioners may now refine intuitions — e.g. that SGD descends on the training loss — to account for noise.

5.1. Future work

Acknowledgments

We thank Sho Yaida, Joshua B. Tenenbaum, Dan A. Roberts, and Greg Wornell for generous mentorship. It was Sho who suggested the question of how to resum; it was Dan who suggested comparing SGD with GD and whose 2018 paper and private notes contains special cases of Corollary FILL IN Roberts (2018). We appreciate writerly feedback from Ben R. Bray, Wenli Zhao, and Chloe Kleinfeldt. We are grateful to our anonymous reviewers' for dramatically improving our writing's clarity.

References

P.L. Bartlett, D.J. Foster, and M.J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 2017.

L. Bottou. Stochastic gradient learning in neural networks. *Neuro-Nîmes*, 1991.

A.-L. Cauchy. Méthode générale pour la résolution des systémes d'équations simultanées. *Comptes rendus de l'Académie des Sciences*, 1847.

P. Chaudhari and S. Soatto. Sgd performs variational inference, converges to limit cycles for deep networks. *ICLR*, 2018.

Laurent Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *ICLR*, 2017.

E. Dyer and G. Gur-Ari. Asymptotics of wide networks from feynman diagrams. *ICML Workshop*, 2019.

C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *ICML*, 2017.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd. *Data @ Scale*, 2018.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better. *NeurIPS*, 2017.

N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P.T.P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *ICLR*, 2017.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 1952.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms i. *PMLR*, 2017.

Qianli Liao, B. Miranda, A. Banburski, J. Hidary, and T. Poggio. A surprising linear relationship predicts test performance in deep networks. *Center for Brains, Minds, and Machines Memo 91*, 2018.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *NeurIPS*, 2017a.

B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro. Geometry of optimization and implicit regularization in deep learning. *Chapter 4 from Intel CRI-CI: Why and When Deep Learning Works Compendium*, 2017b.

H. Robbins and S. Monro. A stochastic approximation method. *Pages 400-407 of The Annals of Mathematical Statistics.*, 1951.

D.A. Roberts. Sgd implicitly regularizes generalization error. *NeurIPS: Integration of Deep Learning Theories Workshop*, 2018.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Berkeley Symposium on Mathematical Probability*, 1956.

Huan Wang, N.S. Keskar, Caiming Xiong, and R. Socher. Identifying generalization properties in neural networks. *Arxiv Preprint*, 2018.

Mingwei Wei and D.J. Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *Arxiv Preprint*, 2019.

P. Werbos. Beyond regression: New tools for prediction and analysis. *Harvard Thesis*, 1974.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning. *NeurIPS*, 2018.

Sho Yaida. Fluctuation-dissipation relations for sgd. *ICLR*, 2019.

Chiyuan Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *ICLR*, 2017.

Appendix A. My Proof of Theorem 1

This is a boring technical proof.

Appendix B. My Proof of Theorem 2

This is a complete version of a proof sketched in the main text.