

論文の内容の要旨

論文題目 Deep Learning Dynamics: Diffusion Perspectives
(深層学習のダイナミクス : 拡散の視点)

氏 名 謝 沢 柯

Abstract:

In recent years, deep learning [1] has achieved great empirical success in various application areas. Due to the highly complex loss landscape, training deep networks is a difficult challenge for deep learning. Stochastic optimization methods, including Stochastic Gradient Descent (SGD) and its variants, are mainstream methods for training deep networks. However, the theoretical mechanism behind stochastic optimization for deep networks is underexplored. In this thesis, we focus on theoretically analyzing deep learning dynamics and designing novel stochastic optimization methods by using a proposed diffusion theoretical framework.

The first part of the thesis investigates how deep learning can select flat minima and escape saddle points. We revealed how minima selection quantitatively depends on the minima sharpness and the hyperparameters. To the best of our knowledge, we are the first to prove that SGD favors flat minima exponentially more than sharp minima, while Gradient Descent with injected Gaussian noise favors flat minima only polynomially more than sharp minima. Adaptive Momentum Estimation (Adam) is the most popular stochastic optimizer for accelerating training of deep neural networks. We prove that Adam can escape saddle points efficiently, but cannot select flat minima as SGD does. This mathematically explains why SGD generalizes better, while Adam generalizes worse but converges faster.

The second part of the thesis investigates how to design better optimization methods and understand other advanced optimization methods. For example, we propose a novel adaptive optimization framework, called Adaptive Inertia, which uses parameter-wise adaptive inertia to accelerate training and provably favors flat minima as well as SGD. We also introduce our recent neuroscience-inspired training algorithm and provide a diffusion interpretation. We believe the proposed diffusion theory will become a powerful theoretical tool for understanding and designing deep learning dynamics in future.

Chapter 1 Introduction.

We introduce the motivation, the background, and the structure of the thesis.

Due to the over-parametrization and the highly complex loss landscape of deep networks, optimizing deep networks is a difficult task. Training deep networks via standard Gradient Descent (GD) can easily get stuck near sharp minima or near saddle points. Empirically, SGD can usually find flat minima that generalize well [2,3]. Adam[4], which combines Adaptive Learning Rate and Momentum, is the most popular stochastic optimizer for accelerating the training of deep neural networks. However, Adam often generalizes worse and finds sharper minima than SGD. The theoretical mechanism behind SGD and Adam is still unclear.

The diffusion theory is an important theoretical tool to understand how deep learning dynamics works [5]. The learning dynamics of deep networks is chaotic. Precisely predicting the evolution of a stochastic dynamical system is a nearly impossible task. This line of research turns to modeling the continuous-time diffusion process of probability densities of parameters instead of modeling parameters themselves.

From the perspectives of both optimization and generalization, stochastic optimization become the dominated optimization method for deep learning. However, existing literature lacks a quantitative theory that answers why stochastic optimization is successful or even necessary for deep learning? Particularly, how can deep learning dynamics select flat minima and escape saddle points well given the complex non-convex loss landscape? The thesis focuses on two fundamental issues in deep learning theory. First, why is deep learning dynamics via stochastic optimization so successful? Second, how to manipulate deep learning dynamics for improving deep learning?

Chapter 2 Recent Advances in Deep Learning Dynamics.

We survey recent advances in deep learning dynamics and discuss the related papers. Understanding deep learning dynamics is critically important for understanding and improving deep learning. Good theoretical analysis can not only explain the current success of deep learning, but also reveal new approaches to future advancements. We mainly discuss recent advances in the following directions, including Langevin Dynamics, Stochastic Gradient Noise (SGN), and Deep Loss Landscape.

Langevin Dynamics [9] is an important theoretical tool for analyzing dynamics governed by potential energy and random noise. Langevin Dynamics-based methods have been proposed for performing Bayesian inference [10]. The Langevin Dynamics approach can be also employed to study continuous-time dynamics of deep learning [11] under stochastic gradient noise which is anisotropic and parameter-dependent. They are two most important applications of Langevin Dynamics in deep learning.

SGN which acts as implicit regularization is considered as an essential advantage for deep learning dynamics [12]. Two lines of research about SGN attract much attention recently: the structure of SGN [13] and manipulating SGN [14]. The three key properties about the structure of SGN are the noise type, the noise magnitude, and the noise covariance. SGN in deep learning is highly position-dependent and anisotropic. Researchers manipulate SGN usually for improving generalization or optimization. Because SGN can effectively regularize deep learning and help escape saddle points and sharp minima.

The loss landscape of deep learning closely relates to the multilayered structure and the overparameterization of deep learning. This line of research focuses on the number of bad local minima [15] and the Hessian the Hessian/curvature [16] of deep loss landscape. One important advantage of deep loss landscape is that, under mild assumptions, deep learning surprisingly has no bad local minima. With the increasing the depth, the number of bad local minima decrease significantly. A mysterious property of deep loss landscape is the ill-conditioned Hessian spectrum. In the context of deep learning, the spectrum of the Hessian is composed of two parts: (1) the bulk centered near zero, (2) and outliers away from the bulk. And, surprisingly, learning mainly happens in a tiny subspace spanned by the eigenvectors corresponding to large eigenvalues of the Hessian.

Chapter 3 Diffusion Theory for SGD Dynamics.

We propose the foundation of the diffusion theory for SGD dynamics [5], and further answer how SGD can select flat minima with a high probability.

Stochastic Gradient Descent (SGD) and its variants are mainstream methods for training deep networks in practice. SGD is known to find a flat minimum that often generalizes well. However, it is mathematically unclear how deep learning can select a flat minimum among so many minima. To answer the question quantitatively, we develop a density diffusion theory to reveal how minima selection quantitatively depends on the minima sharpness and the hyperparameters. The previous papers mainly analyzed the diffusion process under parameter-independent and isotropic gradient noise, while SGN is highly parameter-dependent and anisotropic in deep learning dynamics. Thus, they failed to quantitatively formulate how SGD selects flat minima, which closely depends on the Hessian-dependent structure of SGN. We try to bridge the gap between the qualitative knowledge and the quantitative theory for SGD in the presence of parameter-dependent and anisotropic SGN.

Our theory provides several important insights about deep learning dynamics. First, to the best of our knowledge, we are the first to theoretically and empirically reveal that SGD favors flat minima exponentially more than sharp minima. Second, we explain why large-batch training can easily get trapped near sharp minima, and increasing the learning rate proportionally is helpful for large-batch training. We suggest that either a small learning rate or large-batch training requires exponentially many iterations to escape minima in terms of ratio of batch size and learning rate. Third, although the parameter space is very high-dimensional, SGD dynamics hardly depends on those “meaningless” dimensions with small second order directional derivatives. This novel characteristic of SGD significantly reduces the explorable parameter space around one minimum into a much lower dimensional space. This may explain why learning happens in the relatively low dimensional subspace corresponding to top eigenvalues of the Hessian.

We also directly validate the minima selection formulas on real-world datasets by using neural networks and test functions. Each escape process, from the inside of loss valleys to the outside of loss valleys, are repeatedly simulated for 100 times under various gradient noise scales, batch sizes, learning rates, and sharpness.

Chapter 4 Momentum and Adam Dynamics.

We study the diffusion theory for Momentum and Adam dynamics in terms of saddle-point escaping and minima selection [6]. Adam is the most popular stochastic optimizer for accelerating the training of deep neural networks but often generalizes worse and finds sharper minima than SGD.

Many variants of Adam have been proposed to improve performance. Many of them introduce extra hyperparameters that require tuning effort. Most variants often generalize better than Adam, while they may still not generalize as well as SGD with

tuned hyperparameters. Moreover, they do not have theoretical understanding of minima selection. Loosely speaking, our analysis suggests that Adam variants may also be bad at selecting flat minima due to Adaptive Learning Rate.

Previous work rarely reflect the whole picture of the dynamics of SGD and Adam. Empirically, it does not explain why Adam converges faster than SGD. Moreover, theoretically, all previous works have not touched the saddle-point escaping property of the dynamics, which is considered as an important challenge of efficiently training deep networks

To the best of our knowledge, we are the first to theoretically disentangle the effects of Adaptive Learning Rate and Momentum in terms of saddle-point escaping and minima selection. We quantitatively reveal that, Adaptive Learning Rate is good at escaping saddle point but not good at selecting flat minima, while Momentum helps escape saddle point and matters little to minima selection. Adaptive Learning Rate significantly regularize the Hessian-dependent properties of both saddle-point escaping and minima selection. Momentum provides a drift effect when passing through saddle points, while the drift effect does not change minima selection. Thus, Adam can escape saddle points efficiently, but cannot favor flat minima as well as SGD. This explains why Adam usually converges faster but generalizes worse than SGD.

Chapter 5 Adaptive Inertia Optimizer.

Motivated by the diffusion theory, we try to study how to escape saddle points fast while learning flat minima well. We propose a novel Adaptive Inertia (Adai) optimization framework [6], which is orthogonal to the existing adaptive gradient framework.

Adai does not element-wisely adjust learning rates, but element-wisely adjusts the momentum hyperparameter. Adai combined the advantage of Adaptive Learning Rate and Momentum. Adai can use parameter-wise adaptive inertia to accelerate training and provably favors flat minima well. It can be regarded as an adaptive variant of the classical Heavy Ball Method.

Our theoretical analysis guarantee the convergence of Adai as a stochastic optimizer. We also theoretically prove that Adai can select flat minima as well as SGD. Our empirical results support that Adai significantly outperforms SGD and the Adam variants for deep learning in practice. In numerical experiments, Adai and SGD favor flat minima more than Adam as our theory predicts.

Chapter 6 Neural Variable Risk Minimization.

We introduce our neuroscience-inspired work neural variable risk minimization (NVRM) [7].

Deep learning is often criticized by two serious issues which rarely exist in natural nervous systems: overfitting and catastrophic forgetting. It can even memorize randomly labelled data, which has little knowledge behind the instance-label pairs. When a deep network continually learns over time by accommodating new tasks, it usually quickly overwrites the knowledge learned from previous tasks. Referred to as the neural variability [8,17], it is well-known in neuroscience that human brain reactions exhibit substantial variability even in response to the same stimulus. This mechanism balances accuracy and plasticity/flexibility in the motor learning of natural nervous systems [18].

Thus, it motivates us to design a similar mechanism named artificial neural variability (ANV), which helps artificial neural networks learn some advantages from natural neural networks. We rigorously prove that ANV plays as an implicit regularizer of the mutual information between the training data and the learned model. This result theoretically guarantees ANV a strictly improved generalizability, robustness to label noise, and robustness to catastrophic forgetting [19]. A beautiful coincidence in neuroscience is that neural variability in the rate of response to a steady stimulus also penalizes the information carried by nerve impulses (spikes) [20].

We then devise a NVRM framework and neural variable optimizers to achieve ANV for conventional network architectures in practice. The proposed NVRM framework is an efficient approach to achieving ANV for artificial neural networks. The empirical results demonstrate that our method can (1) enhance the robustness to weight perturbation, (2) improve generalizability, (3) relieve the memorization of noisy labels, and (4) mitigate catastrophic forgetting. Particularly, NVRM, an optimization approach, may handle memorization of noisy labels well at negligible computational and coding costs. One line code of importing a neural variable optimizer is all you need to achieve ANV for your models.

Moreover, we also show that it is useful to theoretically understand the advantage of NVRM from a diffusion perspective.

Chapter 7 Summary.

This thesis provides comprehensive theoretical and empirical analysis of deep learning dynamics from diffusion perspectives. We not only study deep learning dynamics but also propose a diffusion theoretical framework for analyzing deep learning dynamics. The proposed diffusion theory may serve as a powerful theoretical tool for understanding and improving deep learning.

We summarize the main contributions of the thesis in the following. First, we propose a diffusion theory for deep learning dynamics. It reveals the fundamental roles of gradient noise, batch size, the learning rate, and the Hessian in minima selection. Second, by using the diffusion theory, we disentangle the effects of Adaptive Learning Rate and Momentum in Adam learning dynamics and characterize their behaviors in terms of escaping saddle-point and minima selection, which explains why Adam usually converges fast but does not generalize well. Third, we propose a novel optimization framework Adai that elementwisely adjusts the momentum hyperparameter. We show Adai can accelerate training and provably favor flat minima. Fourth, we introduce Neural Variable Risk Minimization and show how to use the diffusion theory as a general theoretical tool to understand its novel learning dynamics.

References:

- [1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444.
- [2] Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural computation*, 9(1), 1-42.
- [3] Hochreiter, S., & Schmidhuber, J. (1995). Simplifying neural nets by discovering flat minima. In *Advances in neural information processing systems* (pp. 529-536).
- [4] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.
- [5] Xie, Z., Sato, I., & Sugiyama, M. (2021). A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast. In *International Conference on Learning Representations*.
- [6] Xie, Z., Wang, X., Zhang, H., Sato, I., & Sugiyama, M. (2020). Adai: Separating the Effects of Adaptive Learning Rate and Momentum Inertia. *arXiv preprint arXiv:2006.15815*.
- [7] Xie, Z., He, F., Fu, S., Sato, I., Tao, D., & Sugiyama, M. (2021). Artificial Neural Variability for Deep Learning: On Overfitting, Noise Memorization, and Catastrophic Forgetting. *Neural computation (to appear)*.
- [8] Churchland, M. M., Byron, M. Y., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., ... & Shenoy, K. V. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature neuroscience*, 13(3), 369-378.
- [9] Coffey, W., & Kalmykov, Y. P. (2012). *The Langevin equation: with applications to stochastic problems in physics, chemistry and electrical engineering* (Vol. 27). World Scientific.
- [10] Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 681-688).
- [11] Li, Q., Tai, C., & Weinan, E. (2019). Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *J. Mach. Learn. Res.*, 20, 40-1.
- [12] Zhu, Z., Wu, J., Yu, B., Wu, L., & Ma, J. (2019, May). The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. In *International Conference on Machine Learning* (pp. 7654-7663). PMLR.
- [13] Simsekli, U., Sagun, L., & Gurbuzbalaban, M. (2019, May). A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning* (pp. 5827-5837). PMLR.
- [14] Xie, Z., Yuan, L., Zhu, Z., & Sugiyama, M. (2021). Positive-Negative Momentum: Manipulating Stochastic Gradient Noise to Improve Generalization. *arXiv preprint arXiv:2103.17182*.
- [15] Kawaguchi, K. (2016, December). Deep learning without poor local minima. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 586-594).
- [16] Sagun, L., Bottou, L., & LeCun, Y. (2016). Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*.
- [17] Dinstein, I., Heeger, D. J., & Behrmann, M. (2015). Neural variability: friend or foe?. *Trends in cognitive sciences*, 19(6), 322-328.
- [18] Fethers, L. (2010). Perspective on variability in the development of human action. *Physical therapy*, 90(12), 1860-1867.
- [19] McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165). Academic Press.
- [20] Stein, R. B., Gossen, E. R., & Jones, K. E. (2005). Neuronal variability: noise or part of the signal?. *Nature Reviews Neuroscience*, 6(5), 389-397.