## **Scientific Progress Report**

## **The Adelis Research Fund for Artificial Intelligence**

**PIs: Yair Carmon, Nadav Cohen, Tomer Koren**

**School of Computer Science, Tel Aviv University**

This report details the scientific progress and key achievements of the research funded by The Adelis Research Fund for Artificial Intelligence at Tel Aviv University, entitled “Interplay between optimization and generalization in modern machine learning,” in the period between 1/10/2021 and 30/9/2022.

## A. Project overview

Optimization algorithms are the backbone of modern machine learning. However, the answer to “what do these algorithms optimize for” remains surprisingly elusive: the ultimate objective of generalization (accurate performance on unseen data) is difficult to analyze, and classical approaches rely on uniform convergence theory to simplify the problem. Unfortunately, uniform convergence demonstrably fails to explain generalization both in practice (e.g. for deep neural networks), and in simple theoretical models (such as stochastic convex optimization). As a result, despite intense research interest, existing theory provides only limited guidance for key decisions in designing optimization methods for machine learning. The goal of the project summarized in this report is to obtain a better understanding of how key degrees of freedom in optimization (such as batch size, learning rate schedule, normalization schemes and properties of the objective function) impact generalization.

## B. Summary of results and achievements

**1. Stability vs Implicit Bias of Gradient Methods**

An influential line of recent work has focused on the generalization properties of unregularized gradient-based learning procedures applied to separable linear classification with exponentially-tailed loss functions. The ability of such methods to generalize well has been attributed to their implicit bias towards large margin predictors, both asymptotically as well as in finite time. We give an additional unified explanation for this generalization and relate it to two simple properties of the optimization objective, that we refer to as realizability and self-boundedness. We introduce a general setting of unconstrained stochastic convex optimization with these properties, and analyze generalization of gradient methods through the lens of algorithmic stability. In this broader setting, we obtain sharp stability bounds for gradient descent and stochastic gradient descent which apply even for a very large number of gradient steps, and use them to derive general generalization bounds for these algorithms. Finally, as direct applications of the general bounds, we return to the setting of linear classification with separable data and establish several novel test loss and test accuracy bounds for gradient descent and stochastic gradient descent for a variety of loss functions with different tail decay rates. In some of these cases, our bounds significantly improve upon the existing generalization error bounds in the literature.

This work is joint with MSc student Matan Schliserman, and has been published as a paper in the 35th Conference on Learning Theory (COLT’22).

**2. Uniform Stability for First-Order Optimization**

We consider the problem of designing uniformly stable first-order optimization algorithms for empirical risk minimization. Uniform stability is often used to obtain generalization error bounds for optimization algorithms, and we are interested in a general approach to achieve it. For Euclidean geometry, we suggest a black-box conversion which, given a smooth optimization algorithm, produces a uniformly stable version of the algorithm while maintaining its convergence rate up to logarithmic factors. Using this reduction we obtain an algorithm for smooth optimization with (nearly) optimal convergence and uniform stability rates, resolving an open problem in the literature. For more general geometries, we develop a novel variant of Mirror Descent for smooth optimization with optimal tradeoff between convergence and stability, leaving open the question of devising a general conversion method as in the Euclidean case.

This work is joint with MSc student Amit Attia, and has been published as a paper in the 35th Conference on Learning Theory (COLT’22).

**3. Continuous vs. Discrete Optimization**

Existing analyses of optimization in deep learning —including its impact on generalization— are either continuous, focusing on (variants of) gradient flow, or discrete, directly treating (variants of) gradient descent. Gradient flow is amenable to theoretical analysis, but is stylized and disregards computational efficiency. The extent to which it represents gradient descent is an open question in the theory of deep learning. The current work studies this question. Viewing gradient descent as an approximate numerical solution to the initial value problem of gradient flow, we find that the degree of approximation depends on the curvature around the gradient flow trajectory. We then show that over deep neural networks with homogeneous activations, gradient flow trajectories enjoy favorable curvature, suggesting they are well approximated by gradient descent. This finding allows us to translate an analysis of gradient flow over deep linear neural networks into a guarantee that gradient descent efficiently converges to global minimum almost surely under random initialization. Experiments suggest that over simple deep neural networks, gradient descent with conventional step size is indeed close to gradient flow. We hypothesize that the theory of gradient flows will unravel mysteries behind optimization and generalization in deep learning.

This work is joint with MSc student Omer Elkabetz, and has appeared as a paper in the 35th Neural Information Processing Systems conference (NeurIPS’21).

**4. Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks**

In the pursuit of explaining implicit regularization in deep learning, prominent focus was given to matrix and tensor factorizations, which correspond to simplified neural networks. It was shown that these models exhibit implicit regularization towards low matrix and tensor ranks, respectively. Drawing closer to practical deep learning, the current work theoretically analyzes the implicit regularization in hierarchical tensor factorization, a model equivalent to certain deep convolutional neural networks. Through a dynamical systems lens, we overcome challenges associated with hierarchy, and establish implicit regularization towards low hierarchical tensor rank. This translates to an implicit regularization towards locality for the associated convolutional networks. Inspired by our theory, we design explicit regularization discouraging locality, and demonstrate its ability to improve performance of modern convolutional networks on non-local tasks, in defiance of conventional wisdom by which architectural changes are needed. Our work highlights the potential of enhancing neural networks via theoretical analysis of their implicit regularization.

This work is joint with students Noam Razin (PhD) and Asaf Maman (MSc), and has appeared as a paper in the 39th International Conference on Machine Learning (ICML’22).

**5. Efficient Distributionally Robust Optimization**

Distributionally robust optimization (DRO) is a promising paradigm for making machine learning more robust to differences between the data on which the model is trained to the data on which it is used. In this work, we develop and analyze new algorithms with improved complexity for two extensively-studied DRO problems: group-structured and f-divergence-based. Our approach relies on an accelerated method that queries a *ball optimization oracle*, i.e., a subroutine that minimizes the objective within a small ball around the query point. Our main contribution is efficient implementations of this oracle for DRO objectives. The resulting algorithms substantially improve over the previous state-of-the-art when the number of training examples is large and the required solution accuracy is high.

This work is joint with MSc student Danielle Hausler, and will appear as a paper in the 36th Neural Information Processing Systems conference (NeurIPS’22).

**6. Making the Monteiro-Svaiter Algorithm Optimal and Practical**

The Monteiro-Svaiter (MS) optimization algorithm is a crucial building block in a growing number of theoretically-efficient methods, including the ball oracle acceleration method mentioned above. However, the practical performance of the MS algorithm leaves much to be desired, mainly because each iteration requires solving an expensive implicit equation. In this work, we develop a new variant of the MS algorithm that removes this expensive step and furthermore requires no manual tuning of algorithmic parameters. From a theoretical perspective, for any p≥2 we improve the complexity of convex optimization with Lipschitz p’th derivative by a logarithmic factor, obtaining the best-possible convergence guarantee. From a practical perspective, our method significantly outperforms the original MS algorithm as well as its existing variants and improvements. An open-source implementation of our algorithm has already been used by subsequent work.

This work is joint with MSc student Danielle Hausler, and will appear as a paper in the 36th Neural Information Processing Systems conference (NeurIPS’22).

## C. Scientific publications

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Bibliographic details** | **Status** | **Link** |
| 1 | Amit Attia, Tomer Koren. Uniform Stability for First-Order Empirical Risk Minimization. In Proceedings of the 35th Conference on Learning Theory (COLT’22), pages 3313-3332, 2022. | Published | [link](https://proceedings.mlr.press/v178/attia22a.html) |
| 2 | Matan Schliserman, Tomer Koren. Stability vs Implicit Bias of Gradient Methods on Separable Data and Beyond. In Proceedings of the 35th Conference on Learning Theory (COLT’22), pages 3380-3394, 2022. | Published | [link](https://proceedings.mlr.press/v178/schliserman22a.html) |
| 3 | Omer Elkabetz and Nadav Cohen. Continuous vs. Discrete Optimization of Deep Neural Networks. In Advances in Neural Information Processing Systems 34 (2021): 4947-4960. | Published | [Link](https://openreview.net/forum?id=iX0TSH45eOd) |
| 4 | Noam Razin, Asaf Maman and Nadav Cohen. Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks. In Proceedings of the 39th International Conference on Machine Learning (ICML’22), 18422-18462. | Published | [Link](https://proceedings.mlr.press/v162/razin22a.html) |
| 5 | Yair Carmon, Danielle Hausler. Distributionally Robust Optimization via Ball Oracle Acceleration. In Advances in Neural Information Processing Systems 35 (2022). | To Appear | [Link](https://openreview.net/forum?id=0ZKyTHwF5V1) |
| 6 | Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, Aaron Sidford. Optimal and Adaptive Monteiro-Svaiter Acceleration. In Advances in Neural Information Processing Systems 35 (2022). | To Appear | [Link](https://openreview.net/forum?id=n3lr7GdcbyD) |

## D. Presentations at scientific conferences

|  |  |
| --- | --- |
| **No.** | **Presentation details** |
| 1 | Uniform Stability for First-Order Empirical Risk Minimization: lecture at COLT’22, July 2022, London, UK |
| 2 | Stability vs Implicit Bias of Gradient Methods on Separable Data and Beyond: lecture at COLT’22, July 2022, London, UK |
| 3 | Continuous vs. Discrete Optimization of Deep Neural Networks: lecture at NeurIPS’21, December 2021, virtual |
| 4 | Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks: lecture at ICML’22, July 2022, Baltimore, MD, USA |

## E. Supported MSc & PhD students

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Name** | **Advisor** | **Degree** |
| 1 | Amit Attia | Tomer Koren | PhD |
| 2 | Matan Schliserman | Tomer Koren | MSc |
| 3 | Noam Razin | Nadav Cohen | PhD |
| 4 | Omer Elkabetz | Nadav Cohen | MSc |
| 5 | Asaf Maman | Nadav Cohen | MSc |
| 6 | Danielle Hausler | Yair Carmon | MSc |