AFRL-AFOSR-UK-TR-2022-0045

OPTiMaL
Optimization for Machine Learning: from Robustness to Regularization

Lorenzo, Rosasco
UNIVERSIT DEGLI STUDI DI GENOVA
VIA BALBI 5
GENOVA, , 16126
IT

05/17/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**	
20220517	Final	**START DATE** 20180924	**END DATE** 20210923

4. TITLE AND SUBTITLE
OPTiMaL
Optimization for Machine Learning: from Robustness to Regularization

5a. CONTRACT NUMBER	**5b. GRANT NUMBER**	**5c. PROGRAM ELEMENT NUMBER**
	FA9550-18-1-7009	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Rosasco Lorenzo

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
UNIVERSIT DEGLI STUDI DI GENOVA VIA BALBI 5 GENOVA 16126 IT	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2022-0045

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The project has contributed to the development of optimal and efficient algorithms for large scale machine learning and their applications, with results in four main directions: 1) design of algorithms with budgeted space complexity; 2) design of algorithms with minimal time cost; 3) design of algorithms able to exploit data geometry; 4) application to the development of efficient AI systems for humanoid robotics and for for model independent new physics searches. The results of the project have led to new theoretical results, new software and new intelligent systems for robotics. In total 10 academic publications resulted from this grant. Research results have been published and presented in the top venues in the field.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		
U	U	U	SAR	3

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER** *(Include area code)*
MARK FRIEND	314-235-6292

Final Report for EOARD Grant FA9550-18-1-7009

"OPTiMaLOptimization for Machine Learning: from Robustness to Regularization"

Date

Name of Principal Investigators (PI and Co-PIs):

- e-mail address : lrosasco@mit.edu
- Institution : MaLGa, Università degli Studi di Genova, Massachusetts Institute of Technology, Istituto Italiano di Tecnologia
- Mailing Address : DIBRIS, Università degli studi di Genova, Via Dodecaneso 35, 16146, Genova, ITALY
- Phone : +39 010 3536607
- Fax : +39 010 3532948; +39 010 3532154

Period of Performance: Sept/Twenty Forth/2018 – September/Twenty Third/2021

Abstract: The project has contributed to the development of optimal and efficient algorithms for large scale machine learning and their applications, with results in four main directions: 1) design of algorithms with budgeted space complexity; 2) design of algorithms with minimal time cost; 3) design of algorithms able to exploit data geometry; 4) application to the development of efficient AI systems for humanoid robotics and for for model independent new physics searches. The results of the project have led to new theoretical results, new software and new intelligent systems for robotics. They have been published and presented in the top venues in the field.

Introduction: While the prospects yield by machine learning (ML) are thrilling, a close look at the human and energetically resources needed by state of the art solutions is worrisome. Software and hardware advances make it easy to follow a brute force approach to ML development, based on deploying more and more resources. This ML growth model poses challenges that might endanger its potential benefits. To tackle the above issues, in this project we developed a novel approach towards efficiency, and hence sustainable machine learning. We rethought the way algorithms are designed and deployed, and we proposed a new multidisciplinary integrated effort, blending statistical and computational aspects for the development of scalable and efficient algorithms.

Results and Discussion:

1. *General iterative regularization.* In a series of recent papers, we started considering the framework of *implicit/iterative regularization*, since it provides a natural way to bridge statistics and optimization. This idea is very popular in practical approaches, but not well very well understood from a theoretical point of view. In particular, the goal of machine learning is to minimize the prediction error on unseen data (test error), while only an approximation of it is available, based on the training data. In this view, the optimization problems that are practically feasible are based on inexact quantities that are stochastic in nature. In [6], we show how probabilistic results, specifically gradient concentration, can be combined with results from inexact optimization to derive sharp test error guarantees. The implicit regularization properties of optimization for learning are highlighted, since we consider unconstrained objective functions. In the same direction, in [4], we started investigating convergence properties of optimization methods which do not use first order information, which are relevant for black box optimization problems, as in reinforcement learning approaches to robotics, where the gradient cannot be computed. More specifically, we proposed and analyze a randomized zeroth-order approach based on approximating the exact gradient by finite differences computed in

a set of orthogonal random directions that changes with each iteration. Our main contribution is proving convergence guarantees as well as convergence rates. Finally, we collected classical and more recent results on implicit regularization in the survey chapter [10].

2. *Acceleration and distributed optimization for learning.* In this task we started from another key question: how classical acceleration schemes impact prediction. Indeed, it is not clear whether driving the training error fast to zero is a desirable property, when minimizing the test error is the goal. In this direction, we studied the convergence behavior of accelerated methods of inertial type, assuming that the objective function satisfies geometrical conditions which are typically satisfied by common loss functions used in machine learning approaches.. In [8], we derived convergence rates for the considered methods in the continuous setting. Further, we considered the regularization effect of distributed computations, which are particularly useful for large-scale problems. In [2], we proposed a new large-scale solver for kernel ridge regression. Our approach combines partitioning with random projections and iterative optimization to reduce space and time complexity while provably maintaining the same statistical accuracy. In particular, constructing suitable partitions directly in the feature space rather than in the input space, we promote orthogonality between the local estimators, thus ensuring that key quantities such as local effective dimension and bias remain under control. We characterize the statistical-computational tradeoff of our model, and demonstrate the effectiveness of our method by numerical experiments on large-scale datasets.

3. *Non convex optimization for learning.* In this task, we did the first steps to extend the above ideas to nonconvex settings. In [1], we investigated the case of neural networks. We characterized the function spaces corresponding to neural networks using the theory of reproducing kernel Banach spaces, opening new ways to understand their properties. In particular, we proved a representer theorem for a wide class of reproducing kernel Banach spaces that admit a suitable integral representation and include one hidden layer neural networks of possibly infinite width. In [5], we studied an algorithm to optimize a black-box function (in particular, nonconvex) using only function evaluations. One natural application of our method is hyperparameters tuning for machine learning, where an explicit form for the objective function is not available. In [3] we focused on compressive learning, an approach which consists in compressing the whole dataset down to a single vector of generalized moments, called the sketch. An approximate solution to the original learning task can then be inferred from this sketch. While previous works focused on data-independent approximation schemes, in our approach we use instead the mean embeddings associated with a Nyström approximation. The latter is data-dependent, i.e. the approximation is adaptive to the dataset to sketch. As a consequence we expect to potentially be able to reach a desired accuracy using a smaller sketch size compared to when using random features. Indeed we observed this behavior experimentally for k-means clustering and Gaussian modeling. Finally, note that some of the papers we described previously indeed contain results which apply to the noncovex case, for instance [4] and [8].

4. *Applications: efficient AI systems for humanoid robotics and for for model independent new physics searches.* Within the project we exploited our theoretical findings in two applications domains: humanoids robots, and new physics searches. Regarding the first problem, in [9] we proposed a system architecture for efficiently addressing the whole-body human-like trajectory generation problem for humanoid robots. The architecture builds upon recent machine-learning methods developed for character animation in computer graphics (CG), and makes it possible to deal with problems which are numerical intractable with classical approaches due to the high dimension. Regarding the application to new physics searches, in [7] we introduced a novel anomaly detection algorithm for model-independent NP searches in high energy

physics. Our approach shows dramatic advantages in efficiency, in terms of both training time and computational resources, compared to similar implementations based on neural networks, with comparable performances.

List of Publications and Significant Collaborations that resulted from your AOARD supported project:

[1] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna, Understanding neural networks with reproducing kernel Banach spaces, arXiv preprint arXiv:2109.09710, (2021).

[2] L. Carratino, S. Vigogna, D. Calandriello, and L. Rosasco, Park: Sound and efficient kernel ridge regression by feature space partitions, Advances in Neural Information Processing Systems, 34 (2021).

[3] A. Chatalic, L. Carratino, E. De Vito, and L. Rosasco, Mean Nystrom embeddings for adaptive compressive learning, arXiv preprint arXiv:2110.10996, (2021).

[4] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa, Zeroth order optimization with orthogonal random directions, arXiv preprint arXiv:2107.03941, (2021).

[5] M. Rando, L. Carratino, S. Villa, and L. Rosasco, Ada-bkb: Scalable gaussian process optimization on continuous domain by adaptive discretization, arXiv preprint arXiv:2106.08598, to appear on Proceedings of AISTATS 2022, (2021).

[6] B. Stankewitz, N. Mücke, and L. Rosasco, From inexact optimization to learning via gradient concentration, arXiv preprint arXiv:2106.05397, to appear on COAP, (2021).

[7] M. Letizia, G. Losapio, M. Rando, G. Grosso, L. Rosasco, Efficient kernel methods for model-independent new physics searches, https://ml4physicalsciences.github.io/2021/files/NeurIPS_ML4PS_2021_146.pdf, (2021)

[8] N. Ginatta, V. Apidopoulus, S. Villa, Convergence rates for the Heavy-ball continuous dynamics for non-convex optimization, under Polyak-Lojasiewicz conditioning, arXiv preprint https://arxiv.org/abs/2107.10123, (2021).

[9] P. M. Viceconte, R. Camoriano, G. Romualdi, D. Ferigo, S. Dafarra, S. Traversaro, G. Oriolo, L. Rosasco, and D. Pucci, Adherent: Learning human-like trajectory generators for whole-body control of humanoid robots, IEEE Robotics and Automation Letters, (2022).

[10] E. De Vito, L. Rosasco, A. Rudi, Regularization: from inverse problems to large-scale machine learning, Harmonic and Applied Analysis, Springer, New York, (2021).