Consistency, Robustness and Sparsity for Learning Algorithms

Konsistenz, Robustheit und Dünnbesetztheit von Lern-Algorithmen

Der Naturwissenschaftlichen Fakultät
der
Friedrich-Alexander-Universität Erlangen-Nürnberg

zur Erlangung des Doktorgrades
Dr. rer. nat.

vorgelegt von
Tim Roith
aus
Amberg

Als Dissertation genehmigt von der Naturwissenschaftlichen Fakultät der
Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung:	—
Vorsitzender des Promotionsorgans:	—
Gutachter*in:	Martin Burger
	Dejan Slepčev
	Franca Hoffmann

Acknowledgement

Coming soon...

Contents

List of Figures

Inhalt und Struktur

Diese Arbeit ist in zwei Hauptteile strukturiert, Part I, welche die Themen und Resultate der Publikation präsentiert und erklärt, welche in Part II erneut abgedruckt sind.

Part I: Exposition	Part II: Prints
Chapter 2: Learning Paradigms	—
Chapter 3: Consistent Semi-Supervised Learning on Sparse Graphs	Chapters P1 and P2
Chapter 4: Robust and Sparse Supervised Learning	Chapters P3 to P5

Einleitung und Motivation

Das Gebiet des maschinellen Lernens entstand in den 1950er Jahren, motiviert durch die Idee, einen Computer Algorithmen und Muster entdecken zu lassen, ohne sie explizit von Hand anordnen zu müssen. Nach der Anfangsphase und mehreren "AI-winters" [SG96] haben zahlreiche wichtige Entwicklungen – z. B. die Wiederentdeckung der "Backpropagation", welche ursprünglich auf [Kel60; Ros+62] zurückgeht und dann in [RHW86] popularisiert wurde, siehe z. B. [Sch22] – zur Relevanz der Lernmethoden beigetragen. Die Fortschritte im Bereich von Computer-Hardware, zusammen mit der Verfügbarkeit groSSer Datenmengen, haben schlieSSlich den Enthusiasmus für maschinelles Lernen der letzten Jahre entfacht. Während "deep" Learning Methoden, d. h. Techniken, die mehrere neuronale Layer verwenden, wie sie ursprünglich in [Ros58] vorgeschlagen wurden, die prominentesten Beispiele sind, gibt es eine ganze Familie von lernbasierten Strategien, welche aktiv in Bereichen wie Computer Vision [Cha+21], Sprachverarbeitung [Khu+23] oder für medizinische Zwecke [She+22] angewendet werden. In dieser Arbeit konzentrieren wir uns hauptsächlich auf datenbasierte Ansätze, angewendet auf Klassifizierungsaufgaben, wobei die konkrete Modalität der gegebenen Daten unsere Strategie bestimmt. Wir konzentrieren uns auf überwachtes Lernen – der Datensatz besteht nur aus Eingabe-Ausgabe-Paaren, d. h., er ist vollständig gelabelt – und halb-überwachtes Lernen – die Daten sind nur teilweise gelabelt.

Beide datenbasierten Methoden waren vor allem in den letzten 20 Jahren sehr erfolgreich. Allerdings weisen die manchmal rein heuristischen Lernstrategien auch gravierende Nachteile auf. Beim überwachten Lernen ist man oft an der Generalisierung eines Klassifizierers interessiert, d. h. wie akkurat ist das Ergebnis auf ungesehenen Eingaben,

die nicht Teil der Trainingsdaten sind. In [GSS14] wurde entdeckt, dass die Ausgaben des Klassifizierers durch kleine, scheinbar unsichtbare Störungen, die als *adversarial attacks* bekannt sind, vollständig verfälscht werden können. Allgemeiner führt uns dieses Phänomen zum Thema *Robustheit* unter Eingabestörungen. Nehmen wir an, dass ein Mensch und eine Maschine eine Eingabe x als vom Typ c einstufen würden. In einer eher vagen, aber anschaulichen Formulierung lautet die wichtigste Implikation, die wir für eine Eingabe $\bar{x}$ erhalten wollen

$$\left.\begin{array}{c}\bar{x} \text{ liegt nahe an } x, \\ \bar{x} \text{ wird von einem Menschen noch als } c \text{ eingestuft}\end{array}\right\} \Rightarrow \text{die Maschine stuft } \bar{x} \text{ als } c \text{ ein.}$$

Neben adversarial Examples gehört dazu auch das Ändern der Auflösung von Bildern, welche die Klassifizierung durch einen Menschen nicht verändern, sofern sie hinreichend klein sind. In jedem Fall zeigt das Vorhandensein dieser Störungen kritische Schwächen der Lernmethoden auf und erfordert ein besseres theoretisches Verständnis der verwendeten Modelle. An dieser Stelle wird die mathematische Grundlage des Fachgebiets relevanter und es kommen Eigenschaften ins Spiel, die über die Klassifizierungsleistung hinausgehen und die in dieser Arbeit diskutiert werden.

Im halb-überwachten Setting betrachten wir Graphbasierte Algorithmen, wie sie ursprünglich in [ZGL03] mit dem Graph Laplace vorgeschlagen wurden. Das Hauptproblem, das wir in dieser Arbeit hervorheben, wurde zuerst in [NSZ09] beobachtet, nämlich dass die Klassifizierungsleistung mit steigender Dimension der Daten deutlich abnimmt. Es stellte sich heraus, dass die mit dem Graph-Laplace erhaltenen Lösungen über den gesamten Datensatz hinweg konstant sind, wenn die Dimension gröSSer als zwei ist, was mit dem Sobolev Einbettungssatz [AF03] in Verbindung gebracht werden kann. Dieses Problem zeigt sich vor allem, wenn die Zahl der unbeschrifteten Datenpunkte gegen unendlich geht, was uns zu der Frage der *Konsistenz* für halb-überwachte Algorithmen führt.

Ein Problem, das für überwachte und halb-überwachte Algorithmen gleichermaSSen gilt, ist der hohe Bedarf an Rechenressourcen. Das Training eines neuronalen Netzes erfordert in der Regel den Einsatz von GPUs über einen langen Zeitraum. Dies macht den Prozess einerseits für weniger leistungsfähige Maschinen oder sogar mobile Geräte undurchführbar und erzeugt andererseits groSSe Mengen an CO_2-Emissionen [Hoe+21]. Für graphbasiertes, halb-überwachtes Lernen müssen zunächst Entfernungen zwischen vielen Datenpunkten berechnet werden, um Kantengewichte zu erhalten, was eine kostspielige Aufgabe ist. AuSSerdem skaliert die Rechenkomplexität verschiedener Probleme auf einem gegebenen Graphen mit der Anzahl der Kanten. Beispielsweise skaliert die Laufzeit von Dijkstras Algorithmus zur Berechnung kürzester Pfade in einem Graphen bereits linear mit der Anzahl der Kanten. In dieser Arbeit ist das Schlüsselwort zur Reduzierung der Rechenlast in beiden Fällen *Dünnbesetztheit*. Das Konzept von dünnbesetzten Matrizen ist tief in der numerischen linearen Algebra verwurzelt [Lan52; GV13] und besteht im Wesentlichen darin, Nullen in einer Matrix auszunutzen, um die Berechnungszeit zu beschleunigen. Bei neuronalen Netzen kann dies dadurch erreicht werden, dass die Gewichtsmatrizen der Layer dünnbesetzt sein müssen. Bei Graphen bedeutet

eine dünnbesetzte Konnektivitätsmatrix einfach, dass nur eine kleine Anzahl an Kanten aktiv ist, was ebenfalls die Rechenkosten reduziert.

Beiträge in dieser Arbeit Anknüpfend an die zuvor genannten Themen befasst sich diese Arbeit mit *Konsistenz, Robustheit* und *Dünnbesetztheit* von überwachten und halb-überwachten Lernalgorithmen.

Für letztere betrachten wir hauptsächlich das sogenannte Lipschitz-Learning [NSZ09], für die wir Konvergenz und Konvergenzraten für diskrete Lösungen zu Lösungen im Kontinuum zeigen, wenn die Anzahl der Datenpunkte gegen unendlich geht. Dabei arbeiten wir mit Annahmen, welche sehr dünnbesetzte und daher rechnerisch attraktive Graphen zulässt.

Bei überwachtem Lernen befassen wir uns mit der Robustheit gegen adversarial Attacks und Auflösungsänderungen. Im ersten Fall schlagen wir einen effizienten Algorithmus vor, der die Lipschitz-Konstante [Lip77] eines neuronalen Netzes bestraft und ein damit robustes Netz trainiert. Im Multiresolution-Setting analysieren wir die Rolle von neuronalen Fourier-Operatoren, wie sie in [Li+20] vorgeschlagen wurden, und ihre Verbindung zu normalen Faltungsoperatoren [Fuk80]. Im Hinblick auf die Rechenkomplexität des Trainings neuronaler Netze schlagen wir einen auf Bregman Iterationen basierenden Algorithmus [Osh+05] vor, der dünnbesetzte Gewichtsmatrizen während des gesamten Trainings ermöglicht. Zusätzliche analysieren wir die Konvergenz der stochastische Adaption der ursprünglichen Bregman Iterationen.

Struktur der Exposition In Chapter 2 stellen wir die Lernparadigmen und Grundbegriffe vor, die in dieser Arbeit verwendet werden. AnschlieSSend stellen wir in Chapter 3 die Themen zur Konsistenz beim halb-überwachten Lernen auf Graphen vor. Nach einer erläuternden Einführung heben wir die Hauptbeiträge von [LIP-I; LIP-II] hervor. Dabei versuchen wir Redundanz zu den Publikationen in Part II zu vermeiden und dennoch einen verständlichen Kontext zu ermöglichen. In Chapter 4 kommentieren wir die Themen zum überwachten Lernen. Nach einer zusätzlichen Einleitung enthält das Kapitel drei Abschnitte, in denen die Arbeiten [FNO; CLIP; BREG-I] einzeln vorgestellt werden. SchlieSSlich werden in Chapter 5 die Inhalte der gesamten Arbeit zusammengefasst und mögliche zukünftige Richtungen aufgezeigt.

Publikationen und Beitragsauflistung

Die folgenden Arbeiten sind Teil dieser Dissertation und werden in Part II erneut abgedruckt.

[LIP-I] T. Roith and L. Bungert. “Continuum limit of Lipschitz learning on graphs.” In: *Foundations of Computational Mathematics* (2022), pp. 1–39.

[LIP-II] L. Bungert, J. Calder, and T. Roith. “Uniform convergence rates for Lipschitz learning on graphs.” In: *IMA Journal of Numerical Analysis* (Sept. 2022).

[CLIP] L. Bungert, R. Raab, T. Roith, L. Schwinn, and D. Tenbrinck. "CLIP: Cheap Lipschitz training of neural networks." In: *Scale Space and Variational Methods in Computer Vision: 8th International Conference, SSVM 2021, Proceedings*. Springer. 2021, pp. 307–319.

[BREG-I] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. "A Bregman learning framework for sparse neural networks." In: *Journal of Machine Learning Research* 23.192 (2022), pp. 1–43.

[FNO] S. Kabri, T. Roith, D. Tenbrinck, and M. Burger. "Resolution-Invariant Image Classification based on Fourier Neural Operators." In: *Scale Space and Variational Methods in Computer Vision: 9th International Conference, SSVM 2023, Proceedings*. Springer. 2023, pp. 307–319.

Die folgenden Preprints sind kein Teil dieser Arbeit, geben aber zusätzliche Einsichten in die behandelten Themen.

[LIP-III] L. Bungert, J. Calder, and T. Roith. *Ratio convergence rates for Euclidean first-passage percolation: Applications to the graph infinity Laplacian*. 2022. arXiv: 2210.09023 [math.PR].

[BREG-II] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. "Neural Architecture Search via Bregman Iterations." In: (2021). arXiv: 2106.02479 [cs.LG].

Im Folgenden führen wir TRs Beiträge zu den oben genannten Publikationen auf.

[LIP-I]: Diese Arbeit baut auf den Erkenntnissen von TRs Masterarbeit auf. Es ist allerdings wichtig anzumerken, dass die Resultate signifikant erweitert wurden und konzeptionell stärker als die der Masterarbeit sind, siehe dazu Abschnitt 3.3 in der Dissertation. TR adaptierte die Kontinuum-Limit-Theorie für den L^∞-Fall, erarbeitet die meisten Beweise und schrieb einen groSSen Teil des Papers. In Zusammenarbeit mit LB, identifizierte er entscheidende Gebiets-Annahmen, welche es erlauben auch mit nicht-konvexen Gebieten zu arbeiten und bewies Konvergenz für angenäherte Randbedingungen.

[LIP-II]: In Zusammenarbeit mit LB, arbeitete TR an den Konvergenzbeweisen, basierenden auf den Ideen von JC. Zusammen mit LB und JC bewies er das Hauptresultat und die verschiedenen Lemmata, die darauf hinführen. Hierbei beschäftigte er sich vor allem mit der Adaption der Theorie für AMLEs auf den Graph-Fall, was das entscheidende Element für die ganze Arbeit ist. Weiterhin, trug er zur Gestaltung und Implementierung der numerischen Experimente, die im Paper durchgeführt wurden bei.

[CLIP]: TR erarbeitete den Algorithmus, der im Paper vorgeschlagen wird, zusammen mit LB, basierend auf dessen Idee. Zusammen mit LS, RR und DT führte er die numerischen Beispiele durch und schrieb groSSe Teile des Quellcodes. Weiterhin schrieb er entscheidende Teile des Papers, wobei DT das Dokument Korrektur lies und klarer formulierte.

[BREG-I]: TR erweiterte LBs Idee, Bregman Iterationen für dünnbesetztes Training einzusetzen, konzipiert durch DT. Zusammen mit MB und LB erarbeitete er die Konvergenzbeweise der stochastischen Bregman Iteration. Hier schlug er auch eine fundierte Initialisierungsstrategie vor. Weiterhin führte er die numerischen Beispiele durch und schrieb den gröSSten Teil des Quellcodes.

[FNO]: Diese Arbeit beruht auf SKs Masterarbeit und verwendet die ursprünglichen Ideen MBs, zu Auflösungsinvarianz mithilfe von FNOs. Im Paper erarbeitete TR die Beweise zur Wohldefiniertheit und Fréchet-Differenzierbarkeit, zusammen mit SK. Er schrieb groSSe Teile des Papers und des Source-Codes, wobei DT bei der Korrektur der publizierten Version mitgeholfen hat. Hierbei führte er die numerischen Studien zusammen mit SK durch.

Preface

This work is structured into two main parts, Part I the presentation and explanation of the topics and results presented in Part II, the peer-reviewed articles.

Part I: Exposition	Part II: Prints
Chapter 2: Learning Paradigms	—
Chapter 3: Consistent Semi-Supervised Learning on Sparse Graphs	Chapters P1 and P2
Chapter 4: Robust and Sparse Supervised Learning	Chapters P3 to P5

Part I consists of five chapters, of which the first two give an introduction and explain the paradigms of *unsupervised*, *semi-supervised* and *supervised* learning. The next two chapters are split up thematically, concerning the topics of semi-supervised and supervised learning, respectively. Here, a short overview provides the necessary framework, allowing us to explain the main contributions. The last chapter presents the conclusion. In Part II the following publications are reprinted:

[LIP-I] T. Roith and L. Bungert. "Continuum limit of Lipschitz learning on graphs." In: *Foundations of Computational Mathematics* (2022), pp. 1–39.

[LIP-II] L. Bungert, J. Calder, and T. Roith. "Uniform convergence rates for Lipschitz learning on graphs." In: *IMA Journal of Numerical Analysis* (Sept. 2022).

[CLIP] L. Bungert, R. Raab, T. Roith, L. Schwinn, and D. Tenbrinck. "CLIP: Cheap Lipschitz training of neural networks." In: *Scale Space and Variational Methods in Computer Vision: 8th International Conference, SSVM 2021, Proceedings*. Springer. 2021, pp. 307–319.

[BREG-I] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. "A Bregman learning framework for sparse neural networks." In: *Journal of Machine Learning Research* 23.192 (2022), pp. 1–43.

[FNO] S. Kabri, T. Roith, D. Tenbrinck, and M. Burger. "Resolution-Invariant Image Classification based on Fourier Neural Operators." In: *Scale Space and Variational Methods in Computer Vision: 9th International Conference, SSVM 2023, Proceedings*. Springer. 2023, pp. 307–319.

The following two works that are not part of this thesis but provide an additional insight.

[LIP-III] L. Bungert, J. Calder, and T. Roith. *Ratio convergence rates for Euclidean first-passage percolation: Applications to the graph infinity Laplacian*. 2022. arXiv: 2210.09023 [math.PR].

[BREG-II] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. "Neural Architecture Search via Bregman Iterations." In: (2021). arXiv: 2106.02479 [cs.LG].

TR's Contribution

Here we list TR's contribution to the publications included in the thesis.

[LIP-I]: This work builds upon the findings in TR's master's thesis [Roi21]. It is however important to note that the results constitute a significant extension and are conceptually stronger than the ones in [Roi21], see Section 3.3. TR adapted the continuum limit framework to the L^∞ case, worked out most of the proofs and wrote a significant part of the paper. In collaboration with LB, he identified the crucial domain assumptions that allow to work on non-convex domains and proved convergence for approximate boundary conditions.

[LIP-II]: In collaboration with LB, TR worked on the convergence proofs building upon the ideas of JC. Together with LB and JC he proved the main convergence result and the various lemmas leading up to it. Here, he was especially concerned with the adaptation of the theory of AMLEs to the graph case, with is a crucial element for the whole work. Furthermore, he contributed to the design and implementation of the numerical examples conducted in the paper.

[CLIP]: TR worked out the main algorithm proposed in the paper together with LB, based on LB's idea. Together with LS, RR and DT he conducted the numerical examples and also wrote large parts of the source code. Furthermore, he wrote significant parts of the paper, where DT proofread and clarified the final document.

[BREG-I]: TR expanded LB's ideas of employing Bregman iteration for sparse training, conceptualized by DT. Together with MB and LB, he worked out the convergence analysis of stochastic Bregman iterations. Here, he also proposed a profound sparse initialization strategy. Furthermore, he conducted the numerical examples and wrote most of the source code.

[FNO]: This work is based on SK's master's thesis, employing the initial ideas of MB for resolution invariance with FNOs. In the paper, TR worked out the proofs for well-definedness and Fréchet-differentiability, together with SK. He wrote large parts of the paper and the source code, where DT helped with proofreading of the published version. Here, he conducted the numerical studies in collaboration with SK.

Part I.

Exposition

Chapter 1
Introduction

The field of *machine learning* emerged in the 1950s [Sam59; Ros58], motivated by the idea of letting a computer discover algorithms and patterns without having to explicitly arrange them by hand. After the initial phase and multiple "AI-winters" [SG96], numerous important developments—e.g., the rediscovery of the backpropagation algorithm, originally due to [Kel60; Ros+62] and then popularized in [RHW86], see, e.g., [Sch22]—contributed to the relevance of learning methods. The advances in computer hardware, together with the availability of large amounts of data, finally allowed the machine learning enthusiasm of recent years to spark. While "deep" learning methods—i.e., techniques involving many stacked neural layers as originally proposed in [Ros58]—are the most prominent examples, there is a whole zoo of learning-based strategies that are actively applied in fields like computer vision [Cha+21], natural language processing [Khu+23] or healthcare [She+22]. In this work, we mainly focus on data-driven approaches, applied to classification tasks, where the concrete modality of the given data determines our approach. Namely, we focus on supervised—the dataset consists only of input-output pairs, i.e., is fully labeled—and semi-supervised—the data is only partially labeled—learning tasks.

For both regimes especially the last 20 years have seen great success of these data-driven methods. However, the sometimes purely heuristic learning strategies also exhibit serious drawbacks. In the supervised setting, one is usually interested in the generalization behavior of a learned classifier, i.e., how good is the performance on unseen inputs which are not part of the given training data. Unfortunately, in [GSS14] it was discovered, that this performance can be completely corrupted, by small, seemingly invisible perturbations known as *adversarial attacks*. More generally, this phenomenon leads us to the issue of *input robustness*. Given some input x, suppose that a human and some machine would classify this input to be of type c. In a rather vague but demonstrative formulation, the key implication we want to obtain for an input $\bar{x}$ is

$$\left.\begin{array}{r}\bar{x} \text{ is close to } x,\\ \bar{x} \text{ is still classified as } c \text{ by a human}\end{array}\right\} \Rightarrow \text{the machine classifies } \bar{x} \text{ as } c.$$

Next to adversarial examples this also includes resolution changes of images, which do not change the classification by a human, if they are reasonably small. In any case, the

existence of these perturbations exhibits critical flaws of learning methods and calls for a better theoretical understanding of the employed models. This is where the mathematical foundation of the field becomes more relevant and properties apart from the classification performance come into play, which are discussed within this thesis.

For the semi-supervised setting, we consider graph-based algorithms as originally proposed in [ZGL03] with the graph Laplacian. The main problem we highlight in this thesis was first observed in [NSZ09], namely that the classification performance deters significantly with increasing dimensionality of the data. In fact, it turned out that solutions obtained by the standard graph Laplacian tend to be constant over the whole dataset, whenever the dimension is larger than two, which can be related to the Sobolev embedding theorem [AF03]. This issue is prevalent in the infinite data limit, where a priori we consider the case, when the amount of unlabeled data points goes to infinity, which leads us to the question of *consistency* for semi-supervised algorithms.

An issue that is shared across the supervised and semi-supervised settings is the high demand for computational resources. Training a neural network usually involves the use of GPUs for long amounts of time. On the one hand, this makes the process infeasible for less powerful machines or even mobile devices and on the other hand, generates questionable amounts of CO_2 emissions [Hoe+21]. For graph-based semi-supervised learning one first needs to compute distances between many data points, to obtain edge weights, which itself is a costly task. Furthermore, the computational complexity of various tasks on a given graph scales with the number of edges. For example, the run time for Dijkstra's algorithm to compute shortest paths on a graph, already scales linearly with the amount of edges [Dij22]. In this thesis, the keyword to reduce the computational load in both cases, is *sparsity*. The concept of sparse matrices routes deeply into the field of numerical linear algebra [Lan52; GV13] and basically consists of exploiting zeros in a matrix to speed up the computation time. For neural networks, this can be incorporated by enforcing the weight matrices of the layers to be sparse. For graphs, sparsity of the connectivity matrix simply means that we have only a small amount of active edges, which also reduces the computational cost.

Contributions in This Work Taking up the previously mentioned subjects, this thesis is concerned with *consistency, robustness* and *sparsity* of supervised and semi-supervised learning algorithms.

For the latter, we mainly consider the so-called Lipschitz learning task [NSZ09] for which we prove convergence and convergence rates for discrete solutions to their continuum counterpart in the infinite data limit. Here, we always work in a framework that allows for very sparse and therefore computationally feasible graphs.

In the supervised regime, we deal with input-robustness w.r.t. adversarial attacks and resolution changes. In the first case, we propose an efficient algorithm, penalizing the Lipschitz constant [Lip77] of a neural network, which trains an adversarially robust network. For the multi-resolution setting, we analyze the role of Fourier neural operators as proposed in [Li+20] and their connection to standard convolutional neural layers [Fuk80]. Concerning the computational complexity of neural network training,

we propose an algorithm based on Bregman iterations [Osh+05] that allows for sparse weight matrices throughout the training. We also provide the convergence analysis for the stochastic adaption of the original Bregman iterations.

Structure of The Exposition In Chapter 2 we introduce the learning paradigms and basic notions used throughout this thesis. We then present the topics on consistency for semi-supervised learning on graphs in Chapter 3. After an explanatory introduction, we highlight the main contributions of [LIP-I; LIP-II]. Here, we try to have as little redundancy to the prints in Part II as possible, while still allowing for an understandable context. In Chapter 4 we comment on the supervised part of this thesis. After an additional introduction, the chapter contains three sections presenting the works [FNO; CLIP; BREG-I] individually. Finally, in Chapter 5 we summarize the contents of the whole thesis and provide possible future directions.