# Chapter 1

# Introduction

"*Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed [177].*"

We can think machine learning as a process of creating models, from example data, that can predict values/classes or discover patterns in data. The real world application areas of machine learning techniques include computer vision, speech recognition, bioinformatics, robotics to mention a few. With the extensive development in the field of information and internet technology and computer hardware, the availability of vast amount of raw data is not a big challenge. The major issue is how to extract useful knowledge from such abundant data. In every application domain, large datasets are available, which require a different class of algorithms. For example, microarray technology provides an enormous number of gene expressions of thousands of genes, millions of similar images can be retrieved from a textual or visual query over the internet, and hundreds of music files can be searched based on a particular rhythm. To analyze such large amount of datasets, it is required to develop new algorithms that can learn efficiently from these massive amounts of data.

In the field of large-scale machine learning generally either the number of data is large, the data is very high dimensional, or there are a large number of classes present in the dataset. In this thesis, we consider the scenario where the number of dimensions is significantly greater than the number of instances. Out of various pillars of convex optimization methods to solve such frameworks, this thesis deals with the fast iterative methods. Popular iterative optimization methods can be classified into two parts, the first-order methods and

the second-order methods, based on the order of the derivative of the objective function which is utilized in each iteration. Earlier class of methods consists of simple iterative schemes but are slow to converge, whereas the later class of methods is fast to converge. However, they require second order derivative information which is very sensitive to numerical errors. Two most basic examples of first-order and second-order methods are the steepest gradient descent and Newton's method, respectively.

In this thesis, the first order iterative methods are considered. The very basic steepest descent method solves the following optimization problem using gradient information,

$$\min_{x\in\mathbb{R}^d} f(x).$$

Under the context of machine learning the function $f(\cdot)$ is a differentiable loss function, which we intend to minimize. In the framework of machine learning, solving the above problem results in an over-fitted solution, which does not perform well with the unseen data. In order to obtain a model that generalizes well, we add a regularizer function $g(\cdot)$ to this problem as follows,

$$\min_{x\in\mathbb{R}^d} \{f(x)+g(x)\}.$$

The function $g$ is used to prevent overfitting of the resulting model so that it can also perform well with the unseen data. The additional constraint we add on the two functions is that we consider two convex functions and we assume that the regularization function may be non-differentiable, which makes the problem unsolvable for the simple steepest gradient descent. One solution is to use the subgradient information of the functions, based on which subgradient descent algorithms were proposed in the early literature. However, the convergence of such methods is very slow. In this thesis, we propose new accelerated algorithms to solve such frameworks with faster convergence. In this process, we explored the mathematical areas of operator splitting methods, fixed-point theory and convex optimization and used new concepts of these fields to design new algorithms.

## 1.1 Literature Review

With the aim to propose new faster first-order methods for solving various regularized machine learning frameworks, we performed extensive literature survey, in the fields of

optimization algorithms that are used in machine learning and the fixed-point theory. The following two subsections give the literature survey in detail.

### 1.1.1 First-order Methods

As discussed earlier, optimization algorithms used in machine learning can be divided into two major categories, the first-order methods, and the second-order methods. The first-order methods use the first-order derivative information of the objective function. Few basic gradient descent algorithms are discussed in [29]. The second-order methods are also the gradient-based methods that utilize the second order information of the objective function (either by computing the Hessian or approximating it). Few examples of second-order methods include Newton's method [141], quasi-Newton's method [214], conjugate gradient [91, 86, 92] and Interior-point method [82]. Although the rate of convergence is very fast, the major issue with the second-order methods is the computation of Hessian matrix, which is both computationally costly, highly space demanding as well as more sensitive to numerical errors. Thus, the first-order algorithms are considered to be suitable for the utilization of optimization in machine learning with high-dimensional datasets. Since our contributions belong to this area, we explored this area in more detail.

A detailed literature review of few first-order algorithms is presented in table 1.1. The algorithms of this class can again be classified in different ways based on various factors, such as convex vs. non-convex functions, smooth vs. non-smooth functions, primal vs. dual optimization problems, constrained vs. non-constrained optimization problems, batch vs. stochastic algorithms, accelerated vs. non-accelerated algorithms, etc. In this table, a short description of few of the important contributions is also given.

TABLE 1.1: Survey of the First order Optimization Schemes

| **Methods** | **Contributions** | **Remarks** |
|---|---|---|
| Batch gradient Descent | Vanilla Gradient Descent [29, 88] | The very basic steepest descent approach for unconstrained minimization of a smooth function. Updates are made by multiplication of learning rate with a gradient of parameters. |

*Continued on next page*

| **Methods** | **Contributions** | **Remarks** |
|---|---|---|
| | Gradient Descent with error [31] | The convergence of the gradient descent algorithm is proved with the notion of an error added in each iteration. |
| | Gradient Descent with Momentum [169, 159] | Few conceptual contributions related to momentum, that is added to the classical gradient descent algorithm for fast convergence. |
| | Line search techniques for gradient descent [7, 88, 29] | Methods discussing the convergence of gradient descent algorithm based on various line-search techniques. |
| Stochastic Gradient Descent | Basic Stochastic gradient descent and its variants [112, 205, 41, 98, 221, 72] | The basic stochastic gradient method for the use of optimization in machine learning tasks is discussed. Variants include various regualrized, composite and dual-averaging methods. Included very basic works in this direction. |
| Subgradient Descent | Basic Subgradient Descent [180, 43, 157] | Solves the unconstrained minimization of a non-smooth function, i.e. $\min_x f(x)$, for non-smooth $f$. |
| | Projected Subgradient Descent [157, 21, 78] | The method consists of generating a sequence $\{x_n\}$, by updating new iterate in the direction opposite to a subgradient of $f$ at $x_n$ and then projecting the resulting vector orthogonally onto a convex set $C$. When the function to be minimized is differentiable, it is equivalent to the steepest descent algorithm. |
| | Few variants and extensions [137, 133] | Few variants of basic subgradient descent, such as the Primal-dual approach combined with subgradient descent, incremental subgradient descent approach etc. |

*Continued on next page*

| **Methods** | **Contributions** | **Remarks** |
|---|---|---|
| Primal-dual first-order method | Basic algorithm and its accelerated variants [48] | First-order primal-dual algorithm for non-smooth convex optimization problems is solved, and its linear rate is proved. Further, accelerated variants are also proposed. |
| | Survey on primal-dual methods [109] | A very good coverage of the basic principles of the primal-dual approaches and their comparison with other numerical algorithms. |
| | Few recent contributions including theoretical results [201, 36] | The first paper presents the batch and randomized approach, for the primal-dual algorithm and a convex-concave saddle point problem representation of convex minimization is solved. The second paper presents a technique to solve the primal problem of finding the zeros of the sum of a maximal monotone operator and the composition of another maximal monotone operator with a linear continuous operator. |
| Conditional Gradient | Basic Frank-Wolfe Method [84, 79] | It is a projection-free method for convex optimization problem with linear convergence rate. |
| | Few variants and extensions of conditional gradient method [111, 104] | Few variants of basic conditional gradient algorithm are proposed, such as away-steps Frank-Wolfe, Pairwise Frank-Wolfe, Fully-Corrective Frank-Wolfe, etc. and their global linear convergence analysis is presented. The second paper presents the primal-dual convergence for few variants of Frank-Wolfe. |

*Continued on next page*

| **Methods** | **Contributions** | **Remarks** |
|---|---|---|
| Projected Gradient Descent | Projected Gradient Descent [29] | A simple and efficient bound-constrained optimization technique, that solves $\min_x f(x)$ subject to $x \in S$, where $S$ is a feasible set. |
| | Few extensions [17, 108] | The first paper discusses the projection methods to solve convex feasibility problem, where the second paper discusses the surrogate projection methods. |
| | Application to Non-negative Matrix Factorization [118] | Practical performance of the method is demonstrated for the task of non-negative matrix factorization. |
| Proximal methods | Monotone theory, General Operator Splitting Methods and Proximal Splitting methods [170, 63, 71, 36, 66, 199, 61, 69, 119, 168, 153, 76] | These papers introduce the notion of proximity operator for solving the convex optimization problems. They also review the basic properties of the proximity operators and other basic and important convergence results. |
| | Douglas-Rachford and ADMM [81, 42, 154] | The papers present the application of Douglas-Rachford splitting and the ADMM (the Douglas-Rachford splitting technique on dual space) on various problems. |
| | Forward-backward Splitting and Proximal gradient Methods and its inexact variant [64, 151, 161, 178, 198] | Explain the forward-backward splitting techniques and the proximal gradient methods in detail. The convergence analysis of proximal gradient algorithm is presented. |

*Continued on next page*

| **Methods** | **Contributions** | **Remarks** |
|---|---|---|
| | Inertial Forward-backward Splitting, Accelerated Proximal gradient Methods, and heavy-ball method [47, 23, 151, 123, 135, 158, 10] | In these papers, the inertial-based forward-backward algorithm and the accelerated proximal gradient algorithms are discussed. |
| | Proximal-average [53, 221, 215, 20] | With the discussion on basic proximal averaging ([20]), to solve the empirical risk minimization problem with the structured properties using the composite penalties, a recent concept of proximal averaging is utilized for different frameworks. |
| Other methods | Bundle-type Methods[113, 115, 181] | The basic Bundle-type methods are discussed and applied to the machine learning, more specifically, to Support Vector estimation, regression, Gaussian Processes, and other regularized risk minimization settings which lead to convex optimization problems. |
| | Mirror-Descent Method [21, 26, 184] | This method solves the minimization of a smooth convex function with a Lipschitz gradient. The key idea in mirror-descent is utilizing a norm, conjugate norm and a distance-generating function in a recursive manner. |

*Continued on next page*

| Methods | Contributions | Remarks |
|---|---|---|
| | Barzilai and Borwein Gradient Method [164] | Solves the large scale unconstrained minimization problem. The main contribution was to redefine the work by [16], which uses second-order information for computing the stepsize. The new work only uses the storage of first-order information during the process. |
| | Coordinate Descent[195] | discussed a coordinate descent algorithm for minimizing the sum of a smooth function and a separable convex function. |
| | Smoothing Techniques [24, 136, 40, 74] | These techniques solve nonsmooth convex minimization problems, constrained or unconstrained, using various smoothing approaches. |
| | Methods for Compressed Sensing/Sparse Recovery [83, 25, 93, 212] | These papers contribute various efficient techniques for sparse recovery/compressed sensing, which directly can be applied to solve a machine learning problem. |

### 1.1.2 Fixed point theory

We not only explored the field of the optimization schemes used in machine learning, but we also surveyed various contributions in the fixed point iterative schemes. The uniqueness of these methods is based on different types of mappings, such as non-expansive, contraction, asymptotically nonexpansive, pseudo-contractive, etc., the design of the iterative scheme, such as two-step, three-step or multi-step and different approaches to mathematical analysis, such as weak convergence or strong convergence, etc. In table 1.2, we present a brief survey. The definitions of all the covered mappings are discussed in appendix. It can be observed in next chapter that the proximal algorithms, where the

proposed algorithms belong, can be interpreted as fixed-point iterative schemes with non-expansive operators. Thus, we give more emphasis to the fixed-point schemes that are specifically designed for this class of mapping.

TABLE 1.2: Survey of the Fixed-point Iterative Schemes

| **Mappings** | **Approaches** | **Contributions** |
|---|---|---|
| Nonexpansive | Mann's Iteration [128, 96, 165, 18, 44] | The very basic Mann's iterations and various weak and strong convergence analysis, at different conditions in Banach and Hilbert spaces are discussed in these papers. |
| | Modified Mann's Iteration [211] | Paper introduces a modified Krasnoselski-Mann iterative algorithm for non-expansive mappings, and strong convergence in Hilbert spaces is proved. |
| | Ishikawa's Iteration [101, 188] | A two step iteration scheme is defined, and the convergence of the scheme is proved. |
| | A three-step Noor's iteration [142, 143, 144] | A new three step iteration by Noor in Hilbert spaces, and its different convergence analysis are discussed. |
| | S-iterations and Normal s-iteration [174, 173] | A new technique is defined using two step fixed-point iterative scheme, which claimed to converge faster than the Picard's and Mann's iteration with contractive mapping setting. |
| | Regularized Methods [114, 208, 176, 175] | Generally result in strong convergence, the regularized methods are inspired from the viscosity-approximation based fixed-point iterative schemes. |

*Continued on next page*

| **Mappings** | **Approaches** | **Contributions** |
|---|---|---|
| | Inertial based Methods [127, 45, 33, 34] | A few recent papers based on analysing the iterative schemes with the inertial effects. |
| Contraction | Method of successive approximations and Banach fixed-point theorem [156, 120, 15] | In [120], the author introduced the method of successive approximations. Picard in [156] developed it systematically and gave a well known proof of existence and uniqueness of the solution of initial value problems for ordinary differential equations. [15] discusses the popular contraction mapping theorem. |
| | Few generalizations of contraction mapping theorem [59, 187, 80] | These paper discuss few generalizations of the contraction mapping theorem under different settings. |
| | New definitions and theorems of contraction mapping [130, 150, 166] | These paper discuss new definitions of the contraction mapping. New theorems under such mappings are also given. |
| | Multi-valued contraction mappings [131, 67, 75] | These works introduce the notion of multi-valued contraction mappings, and various contributions on such mappings. |
| asymptotically-nonexpansive | The s-iteration scheme [1] | A new iterative scheme is proposed. Authors claimed that the rate of convergence of the new scheme is similar to the Picard's iteration and faster than the other fixed-point iteration process. |
| | A three step iterative scheme[206, 57, 185] | The first paper defines and analyses a three-step iterative scheme for asymptotically nonexpansive mappings in Banach spaces. This work is further extended and analyzed in the second and third paper. |

*Continued on next page*

| Mappings | Approaches | Contributions |
|---|---|---|
| | Other effective schemes [179, 28] | This paper [179] and the book [28] explain other effective iterative schemes. |
| pseudo-contractive | Ishikawa's Iterations [100] | Ishikawa initially proposed his new two-step iterative scheme for pseudo-contractive setting and then generalized for nonexpansive in [101]. |
| | Weak and Strong convergence analysis by Mann's Iterations [129, 54] | These methods describe new weak and strong convergence analysis of Mann's iterative schemes. New examples of these schemes are also discussed. |
| | Modified Mann's scheme for not necessarily Lipschitzian [172] | It is proved that the modified Mann iteration process converges weakly to a fixed point of an asymptotically pseudocontractive mapping in the intermediate sense which is not necessarily Lipschitzian. |
| | Other effective schemes [55, 28] | This paper [55] and the book [28] explain other effective iterative schemes. |

## 1.2 Motivation

After extensive initial literature survey, it was found that most of the proximal algorithms that were applied to the machine learning problems were based on the Picard's and Mann's fixed-point iterations [194, 23, 138, 22, 145, 62]. These algorithms are applied to solve various real-world problems, and the practical performance of both the Picard's and Mann's fixed-point iterations are analyzed in depth. In literature, different extensions are also available such as extensions in dual spaces and the stochastic definitions. However, beyond these two iterative schemes, many unexplored fixed-point iterative schemes also exist, which have never been analyzed for solving the machine learning problems using convex optimization.

In addition to the area of proximal algorithms for the task of machine learning, we also explored the current developments and research directions in the area of fixed-point theory

with nonexpansive and contraction mappings. The field of Operator theory is considered to be a part of pure mathematics, which is extremely rich in the novel mathematical concepts and theoretical breakthroughs. The techniques of operator theory applied to solve numerous problems that belong to the field of variational inequality and convex optimization. Although rich in analytical theories, the lack of empirical performance analysis of these techniques is a big issue. There exist lots of unexplored schemes in this area, which are claimed to be faster than the Picard's iterative scheme, however, due to lack of implementation details could not explore practically for the real world problems.

A traditional approach to analyze an algorithm is to prove the convergence of the algorithm. It is observed that for the accelerated gradient algorithms we found in the literature, only weak convergence analysis is given in the general infinite-dimensional Hilbert space. However, the empirical analysis of the practical behaviour of an algorithm which converges strongly in the general infinite dimensional Hilbert space is missing.

In this thesis, we tried to fill these research gaps. We go beyond the limitation of using the traditional fixed-point iterative schemes to define proximal algorithms for machine learning and proposed new definitions of proximal algorithms based on new fixed-point iterative schemes. Few recently proposed fixed-point iterative schemes are exploited by applying them to solve multiple machine learning problems using the lasso framework and its various extensions. Our proposed models analyse few basic but strong concepts of pure mathematics (fixed-point theory) such as viscosity-approximation based fixed-point schemes and extragradient-based fixed-point schemes and their inertial-based variants.

## 1.3 Research Contributions

The research contributions of this thesis are as follows:

- In this work, we first studied the relationship between the proximal methods and the concept of fixed-point iterative schemes and investigated the gap between the lack of applications in ongoing research in the field of fixed-point theory and the requirement of new faster algorithms in the field of first order methods in machine learning. We found that most of the proximal algorithms that are applied to the machine learning employ the very basic Picard and Mann fixed-point iterative schemes.

However, many advanced fixed-point schemes are proposed that are not analyzed for such tasks.

- We designed three advanced fixed-point iterations based proximal algorithms for solving real-world problems of machine learning. A new accelerated gradient algorithm is proposed named as **NAGA**, which is based on the proposed extragradient based forward-backward splitting techniques along with the inertial step, named as (**NIFBA**).

- The convergence of NAGA is analyzed with both the contraction and non-expansive mappings (definitions given in Appendix). We also analyzed the stability of the proposed fixed-point iterative scheme for NAGA, with respect to the contraction mappings. It is shown that NAGA weakly converges to a fixed solution point and is $T$-stable with respect to the contraction operator $T$.

- The newly designed algorithms and the proposed algorithm NAGA are applied to solve various real world problems using the lasso and extended lasso frameworks. The algorithms are applied to (i) the high-dimensional regression problem, (ii) the unified sparse representation learning problem for cross-modal datasets and (iii) the cancer prediction problem with the help of two complex non-smooth penalties for lasso. After performing extensive experimentation with real benchmark datasets, we found that NAGA outperforms in solving all the problems in terms of the number of iterations required to converge, empirical convergence rate and accuracy.

- To analyze the performance of an algorithm that converges strongly to a fixed-point in infinite-dimensional real Hilbert spaces, we designed the latest viscosity-approximation based fixed-point scheme as a viscosity-approximation-based proximal gradient algorithm (**VPGA**). Also, a novel viscosity-approximation based accelerated gradient algorithm (**VAGA**) is also proposed, which is based on proposed viscosity-based inertial forward-backward algorithm (**VIFBA**).

- It is shown that the sequence generated by VAGA is bounded and converges strongly to a fixed-point under specific conditions in infinite-dimensional Hilbert spaces.

- We applied both the VPGA and VAGA algorithms to solve the regularized multitask learning problem that employs multitask lasso framework. The lasso framework consists of the squared loss function along with the $\ell_{2,1}$ norm. Experimental results

on three benchmark multitask regression datasets are presented. With the help of experiments, it is shown that algorithms with the strong convergence also perform well.

- To show the applicability of the proposed algorithms, we also applied the VPGA and VAGA algorithms to the problem of joint splice-site recognition problem of bioinformatics, which utilises the multitask learning framework. Here, seven publicly available gnome datasets are used to recognize the splice-sites in the gnome sequences. We found that solving the multitask framework with the help of viscosity-approximation based proximal algorithms is not only faster, but also these algorithms give a stable solution in comparison to the traditional proximal algorithms.

- In more general settings of operator splitting algorithms, we proposed an extragradient-based operator splitting algorithm (**EOSA**) and its accelerated variant (**AEOSA**) for the problem of finding zeros of the sum of two convex nonsmooth functions. The proposed approach is a natural generalization of the splitting methods of the Peaceman-Rachford operator splitting [153] and the Douglas-Rachford operator splitting [76]. The convergence of both of the algorithms is analyzed.

- Both the EOSA and AEOSA algorithms are applied to solve the lasso problem for the task of microarray gene analysis. Four publicly available real gene-expression datasets are utilized for this task using the lasso framework. From the experimental results, it is found that the newly proposed techniques solve the lasso problem faster than the traditional frameworks.

## 1.4 Layout of Thesis

The rest of thesis is organized as follows. New definitions of traditional proximal gradient algorithms based on different fixed-point iterative schemes and the proposed new accelerated gradient algorithm (**NAGA**) are discussed in Chapter 3. In Chapter 4, we discuss our work on viscosity-approximation based proximal gradient (**VPGA**) and accelerated gradient algorithms (**VAGA**). Chapter 5 discusses our work on the extragradient-based operator splitting algorithm (**EOSA**) and its accelerated variant (**AEOSA**) for the problem of finding zeros of the sum of two convex nonsmooth functions. We present the

concluding remarks and scope of the future research in the last Chapter 6. We begin with presenting few background works and related mathematical concepts in the next Chapter 2.

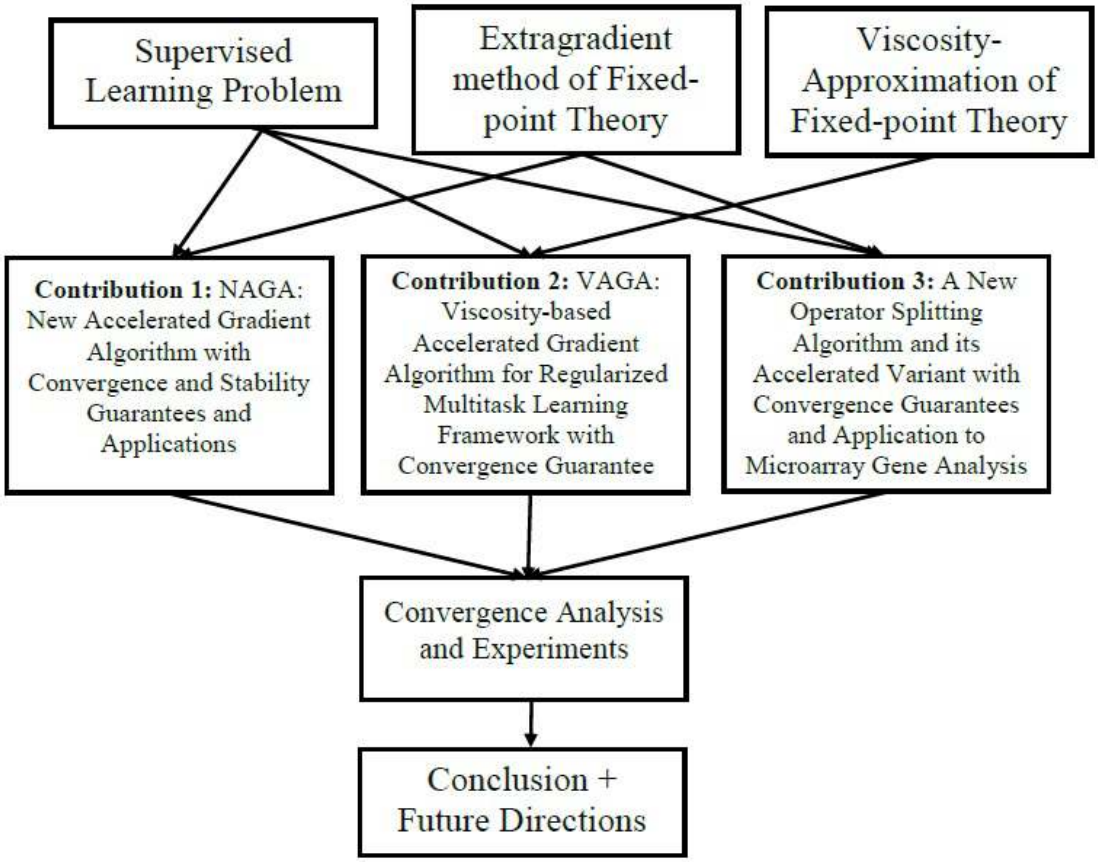


FIGURE 1.1: Graphical Abstract