
ACADEMIA ROMÂNĂ
SCOSAAR

# ABSTRACT OF HABILITATION THESIS

## Stochastic Optimization

Habilitation field: Mathematics

**Ion Necoară**

# Abstract

## 0.1 Contributions of this thesis

Consider the Euclidean space $\mathbb{R}^n$ endowed with the usual scalar product $\langle x, y\rangle = x^T y$ and the corresponding norm $\|x\| = \sqrt{\langle x, x\rangle}$[1]. Optimization problems from this thesis are variants of the following stochastic problem with the objective function expressed in the composite form:

$$\min_{x \in \text{dom}F} F(x) := \mathbf{E}\left[f(x, \xi) + g(x, \xi)\right], \tag{1}$$

where $\xi$ is a random variable over a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Moreover, the functions $f : \mathbb{R}^n \times \Omega \to \bar{\mathbb{R}}$ and $g : \mathbb{R}^n \times \Omega \to \bar{\mathbb{R}}$ are proper, lower semicontinuous, convex in the first argument. The randomness in most of the practical optimization applications led the stochastic optimization field to become an essential tool for many applied mathematics areas, such as machine learning and statistics, control and signal processing, sensor networks and others. The main issue with the stochastic problem (1) is that we cannot evaluate the (sub)gradient $\nabla F(x)$ of function $F$ in a point $x \in \text{dom}F$, that is we cannot have access to the subdifferential $\partial F(x)$. On the other hand, for a given sample $\hat{\xi}$ of random variable $\xi$, we can easily compute $f(x, \hat{\xi}), g(x, \hat{\xi})$ and the (sub)gradients $\nabla f(x, \hat{\xi}), \nabla g(x, \hat{\xi})$, respectively. In this thesis we assume that the (sub)gradients $\nabla f(x, \hat{\xi}), \nabla g(x, \hat{\xi})$ are unbiased estimators, i.e.:

$$\mathbf{E}\left[\nabla f(x, \xi) + \nabla g(x, \xi)\right] \in \partial F(x).$$

Note that the optimization problem (1) is very general and covers many applications from engineering, statistics and machine learning.

A particular case of stochastic problem (1) is $g \equiv 0$ and thus $F(x) = \mathbf{E}\left[f(x, \xi)\right]$. This is the typical problem arising in machine learning applications. In this case, the most used algorithm for solving such a problem is the stochastic gradient descent (SGD) proposed for the first time by Robbins and Monro[2]:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k, \xi_k),$$

where $\xi_k$ is a sample of the random variable $\xi$ (i.e., $\xi_k \sim \mathbf{P}$) and in order to ensure convergence of this stochastic iterative process the stepsize $\alpha_k$ must be chosen as

$$\alpha_k = \frac{\alpha_0}{k^\gamma}, \quad \text{where } \alpha_0 > 0 \text{ and } \gamma \in [0, 1].$$

In general, SGD has slow convergence and thus one of the main research directions in this field, especially in the present era of Big Data, is designing accelerated variants of this algorithm.

[1] Most of the results of this thesis can be easily extended to more general spaces.

[2] H. Robbins and S. Monro, A stochastic approximation method, The Annals of Mathematical Statistics, 1951.

Experimentally, it has been observed that minibatch stochastic gradient descent performs better:

$$x_{k+1} = x_k - \alpha_k \frac{1}{|J_k|} \sum_{\xi \in J_k} \nabla f(x_k, \xi),$$

where the minibatch sample $J_k \subset \Omega$ has cardinality $|J_k| > 1$. However, minibatch SGD may exhibit speedup saturation beyond a particular batchsize, as one can notice from Figure 1[3]. In particular, for large batches we may need a larger number of passes over the dataset, resulting in overall slower computation. Hence, there is an open problem in stochastic optimization to explain when minibatching works. More precisely, one needs to explain mathematically why and when minibatching works in SGD and also to identify the optimal minibach size.

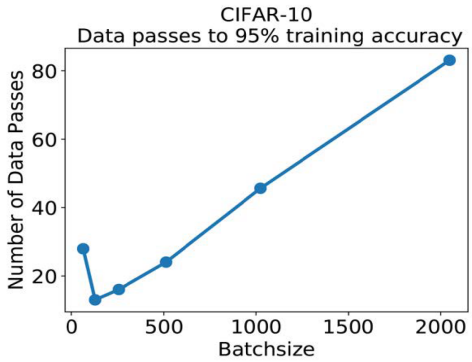


*Figure 1: Behavior of SGD as a function of minibatch size.*

In this thesis we answer (partially) to this yet unsolved problem, deriving conditions when minibatching works for Kaczmarz algorithm in Chapter 3 (recall that Kaczmarz coincides with SGD when one considers the linear least-squares problem), we continue with alternating projection methods for convex feasibility problems in Chapter 4 and Chapter 5 and finally with a stochastic subgradient with alternating projections algorithm for solving optimization problems with many functional constraints in Chapter 7.

In particular, in Chapter 3 we prove that a stochastic minibatch Kaczmarz algorithm, that uses at each step a random subset of the constraints and extrapolated stepsizes, has linear convergence, with a rate depending on the geometric properties of the matrix and its submatrices and on the size of the blocks. Our convergence analysis reveals that the algorithm is most effective when it is given a good sampling of the rows into well-conditioned submatrices. Besides providing a general framework for the design and analysis of stochastic block Kaczmarz methods, our results resolve an open problem in the literature related to the theoretical understanding of observed practical efficiency of extrapolated block Kaczmarz methods. Our framework allows to also identify the optimal minibatch size.

In Chapter 4 we present a family of stochastic projection methods for solving the convex feasibility problem with (possibly) infinite intersection of sets. We prove that under a stochastic linear regularity condition, the algorithms converge linearly, with a rate that has a natural interpretation as a condition number of the stochastic optimization reformulation of the convex feasibility problem and that depends explicitly on the number of sets sampled. This condition number depends on the linear regularity constant and an additional key constant which can be interpreted as a Lipschitz constant of the gradient of the stochastic optimization reformulation. We have identified

[3]Courtesy of Yin et al. 2018.

the Lipschitz constant as the key quantity determining whether extrapolation helps or not, and how much. In Chapter 5, we extend these results to convex feasibility problems, where each set from the intersection is specified algebraically as a convex inequality, where the associated convex function is general (possibly non-differentiable). In this case, the algorithm does not require computation of projections but subgradient updates. For these minibatch stochastic subgradient-based projection methods we also derive sufficient conditions under which the convergence rates depend explicitly on the minibatch size. To the best of our knowledge, these works are the first deriving conditions that show theoretically when minibatch stochastic projection updates have a better complexity than their single-sample variants.

In Chapter 7 we consider convex optimization problems with (possibly) infinite intersection of constraints, each one given as the level set of a convex but not necessarily differentiable function. For these settings we propose stochastic subgradient algorithms where we first take a subgradient step aimed at only minimizing the objective function and then a subsequent subgradient step minimizing the feasibility violation of the observed minibatch of constraints. For extrapolated stepsizes, we prove linear convergence rates that depend explicitly on the minibatch size and show when minibatching helps a subgradient scheme with random feasibility updates.

On the other hand, the convergence theory for SGD has been derived for simple stochastic optimization models satisfying restrictive assumptions, the rates are in general sublinear and hold only for specific decreasing stepsizes. For example, the convergence theory treats separately smooth or non-smooth objective functions, although the convergence rates are the same for these two cases, and covers usually unconstrained optimization models, i.e. $g \equiv 0$. However, in many applications we have regularization terms or constraints which lead to the optimization problem (1), i.e. $g \not\equiv 0$. Convergence analysis of SGD to more general problems, e.g. optimization problem (1), has not been given yet. Therefore, in the second part of this thesis we extend the convergence analysis of stochastic first order methods to the more general problem (1).

First, we extend the convergence analysis of SGD to optimization problems with many (possibly infinite) constraints, where each constraint is expressed either through a convex set (Chapter 6) or through a convex function (Chapter 7). In this case, each individual function $g$ in (1) denotes the indicator function of one set from the intersection defining the feasible set. Moreover, although SGD has cheap iteration and its practical performance may be satisfactory under certain circumstances, there is recent evidence of its convergence difficulties and instability for unappropriate choice of parameters. To avoid some of these drawbacks of SGD, we consider a stochastic proximal point (SPP) algorithm, which is more robust w.r.t. parameters, see Chapter 6. For this method we derive sublinear convergence rates when the objective function is convex or strongly convex, smooth or nonsmooth.

In Chapter 8, we present a general framework for the convergence analysis of stochastic first order algorithms (SGD and SPP) which is based on the assumptions that the objective function satisfies a stochastic bounded gradient condition, with or without a quadratic functional growth property. These conditions include the most well-known classes of objective functions analyzed in the literature: nonsmooth Lipschitz functions and composition of a (potentially) nonsmooth function and a smooth function, with or without strong convexity. Based on this framework we derive a common convergence analysis for these stochastic first order methods. Moreover, our convergence rates are optimal for the classes of problems we consider.

In the last chapter we present some research directions that we will consider in our future work.

Finally, note that all the algorithms from this thesis have been tested numerically on concrete applications, using synthetic or real data, and compared with other state of the art methods from

literature. The numerical results either confirm the theoretical ones or they show the superior performance of our methods.

## 0.2 Articles in ISI journals

The material that is presented in this thesis has been published, or accepted for publication, in top journals. We detail below the main publications from this thesis published in the last 3 years.

- **I. Necoara**, *Faster randomized block Kaczmarz algorithms*, Siam Journal on Matrix Analysis and Applications, vol. 40, nr. 4, 1425–1452, 2019 (Q1 if/ais - Applied Mathematics).
- **I. Necoara**, P. Richtarik, A. Patrascu, *Randomized projection methods for convex feasibility problems: Conditioning and convergence rates*, Siam Journal on Optimization, vol. 29, nr. 4, 2814–2852, 2019 (Q1 if/ais - Applied Mathematics).
- **I. Necoara**, A. Nedich, *Minibatch stochastic subgradient-based projection algorithms for solving convex inequalities*, partially accepted in Computational Optimization and Applications, 2020 (Q1 if/ais - Applied Mathematics).
- A. Nedich, **I. Necoara**, *Random minibatch subgradient algorithms for convex problems with functional constraints*, Applied Mathematics and Optimization, vol. 80, nr. 3, 801–833, 2019 (Q1 if/ais - Applied Mathematics).
- A. Patrascu, **I. Necoara**, *Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization*, Journal of Machine Learning Research, vol. 18, no. 198, 1–42, 2018 (Q1 if/ais - Automation & Control Systems).
- **I. Necoara**, *General convergence analysis of stochastic first order methods for composite optimization*, Journal of Optimization Theory and Applications, doi: 10.1007/s10957-021-01821-2, 2021 (Q2 if/ais - Applied Mathematics).

In two papers from the above list I am the single author, other two are joint works with a former phd student (A. Patrascu) and other two papers are joint collaborations with prof. A. Nedich from Arizona State University (USA).