AFRL-AFOSR-UK-TR-2022-0002

Towards a theory of long-step algorithms for large scale optimization

Bolte, Jerome
Fondat J J Laffont Tlse Sciences Eco
21, Allee De Brienne
TOULOUSE, , 31000
FR

10/29/2021
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

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
29-10-2021	Final	15 Jun 2018 - 14 Jun 2021

4. TITLE AND SUBTITLE
Towards a theory of long-step algorithms for large scale optimization

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-18-1-0226

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Jerome Bolte

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fondat J J Laffont Tlse Sciences Eco
21, Allee De Brienne
TOULOUSE, 31000
FR

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
EOARD
UNIT 4515
APO AE 09421-4515

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-UK-TR-2022-0002

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This research achieved its goal to study the design of large steps first-order methods by exploiting the geometry or the regularity of the problems. New algorithms for optimization were developed, complexity analysis were completed and new geometries were explored for the previously developed NoLips algorithm, an instance of the Bregman method, which has been exploited by the optimization community. This research generated 15 articles and has already garnered more than 250 Google Scholar citations. The final report attached references section has links to the produced articles which provide additional details beyond the overview contained within the final report.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			MARK FRIEND
U	U	U	SAR	10	**19b. TELEPHONE NUMBER** *(Include area code)* 314-235-6292

Final report for grant FA9550-18-1-0226, 2019-2021
"Towards a theory of long-step algorithms for large scale Optimization"

PI : Jérôme Bolte, TSE / Université Toulouse I Capitole.

Abstract. From an "artificial intelligence" perspective, our goal was to find ways and settings in which learning rates may be increased, facilitating therefore training and intelligence acquisition processes. In the optimization terminology, we focused on "large steps" first-order algorithms, both deterministic and stochastic, and designed them by means of various structural and geometrical considerations. In particular, we introduced and studied a notion of relative smoothness for non Euclidean problem which covers a wide range of nonsmooth nonconvex problems. Our basic algorithm called NoLips, an instance of the Bregman method, has already been largely exploited by the optimization community. We provided a surprising optimality result: the optimal complexity of Bregman-like methods is $O(1/k)$ and not $O(1/k^2)$ as in the Euclidean case. We also provided several counterexamples in the *convex* world related to longstanding problems involving large steps methods: the steepest descent method with optimal step does not converge, NoLips does not converge either. We also obtained results in the stochastic world by providing methods with agressive step-sizes: NoLips with variance reduction, SGD without replacement sampling, INNA anmethod for deep learning, second-order stepsize tuning in Deep Learning .

1 Research overview

The goal of this proposal is to study the possibility of designing large steps first-order methods by exploiting the geometry or the regularity of the problems. Making large steps is fundamental for providing efficient and fast methods, either deterministic or stochastic. For instance, in machine learning, large steps amount to high learning rates and result in much faster training.

Our research covers nonsmooth nonconvex, deterministic and stochastic optimization.

The application domains that we considered range from inverse problems in statistics to modern problems in AI as Deep Learning .

The COVID19 crisis compromised considerably the work flow for this proposal. A no-cost extension was awarded.

Research Lines Our work can be divided and understood through different angles.

- **New algorithms for optimization**: we designed methods starting from a given class of problems known for their importance in some field, as for instance nonsmooth problems, composite methods or Deep Learning .
  Applications considered: Nonlinear programming, Deep Learning .

- **New geometries for long steps method**: the ideas behind new geometries are at the origin of the proposal. They were triggered by one of our previous proposals FA9550-14-1-0056, and by the introduction of the NoLips algorithm. These geometries generalize the Euclidean case and they can be thus adapted to match many new problems: either because of the shape of the functions or of the nature of constraints. The question of determining the best steps was a key question in this proposal too.
  Applications considered: Symmetric Nonnegative Factorization, Euclidean Matrix reconstruction.

- **Complexity results**: we tried to analyze upper and lower-bound for complexity for Bregman like methods, composite methods and some minmax problems. In particular we provided the optimal complexity of deterministic Bregman methods.
  Consequences: new guarantees, new paths for acceleration

- **Negative results**: in our quest to augment step-length, we also discovered several limitations that were unknown to the community. Indeed, some well known methods may have an acceptable complexity, yet the underlying sequence may oscillate indefinitely near the minimizers' set.
  Consequences: many counterexamples, search for new rigidity assumptions.

The project in numbers

– 15 articles were produced, some of them appeared in high ranked journals or conferences as NeurIPS, Math. Prog., Math. OR, JEMS, ICML.

– a dozen of conferences, organization of seminars,

– we cofounded the French Optimization Seminar (2020) (in order to foster research in our environment during the COVID crisis)
`https://gdrmoa.math.cnrs.fr/seminaire-francais-optimisation/`

– one prize was obtained, U. Rothblum OR prize, Israel, on the very subject of the proposal:
J. Bolte, S. Sabach, M. Teboulle, Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*

– J. Bolte was awarded a chair in AI within the Artificial and Natural Intelligence Toulouse Institute (ANITI) within Villani's national AI plan.
His chair is entitled "Large scale Optimization for AI"

Involved researchers

– Jérôme Bolte, Full Professor, TSE, ANITI, Université Toulouse Capitole.

– Edouard Pauwels, Assistant Professor, ANITI and University Toulouse 3.

– Radu Dragomir, University Toulouse 3 & ENS Paris, now UC Louvain

– Rodolfo Rios-Zertuche, ANITI and LAAS

2 Main results

2.1 The NoLips algorithm and its extensions

The central method we consider is based on the *Bregman gradient method*

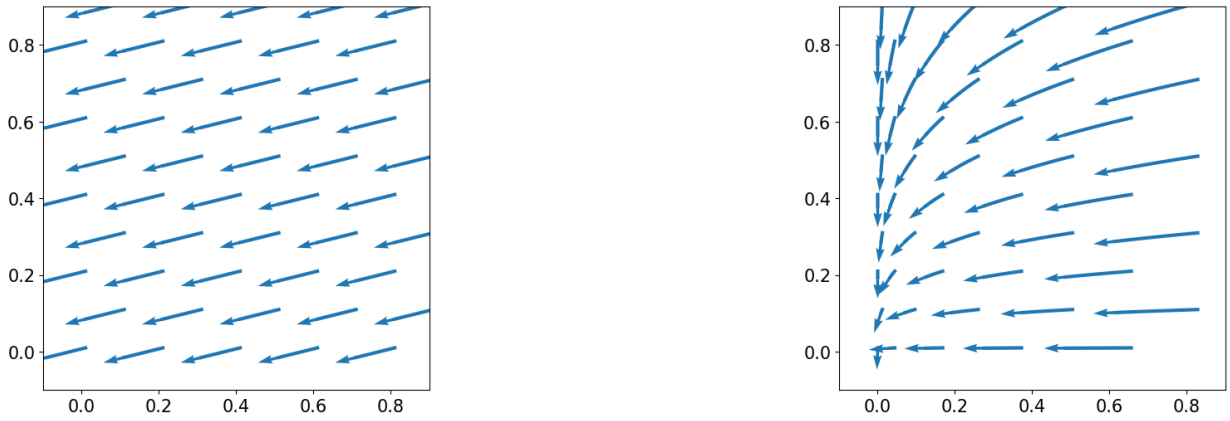
$$x_{k+1} = \arg\min_{u \in C} f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \lambda D_h(u, x_k), \tag{BG}$$

where the Euclidean distance has been replaced by the *Bregman distance* (see Figure 1)

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

induced by some strictly convex and continuously differentiable *kernel function* h. Under the assumption that $Lh - f$ is convex and $\lambda < 1/L$, the complexity of the method is $O(1/k)$. This is the discovery in reference [A] funded by USAF in a previous proposal which has now more than 200 Google Scholar citations. We called the fundamental property "$Lh - f$ convex" relative smoothness.

We focus here on the most salient novelties



(a) Euclidean (b) Entropy

Figure 1: Effect of Bregman divergence on elementary steps

NoLips in the nonconvex world [6] We focused on nonconvex and nonsmooth minimization problems with a composite objective, where the differentiable part of the objective is freed from the usual and restrictive global Lipschitz gradient continuity assumption as explained above. This longstanding smoothness restriction is pervasive in first order methods, and recently was circumvented for convex composite optimization in [A] through a simple framework which captures, all at once, the geometry of the function and of the feasible set. Building on this work, we tackled genuine nonconvex problems. We complemented and extended the approach in [A] to derive an extended descent lemma. We then considered a Bregman-based proximal gradient method for the nonconvex composite model with relatively smooth functions, which is proven to globally converge to a critical point under natural assumptions on the problem's data, and in particular for semialgebraic problems. To illustrate the potential of our general framework and results, we consider a broad class of quadratic inverse problems with sparsity constraints which arises in many fundamental applications, as phase retrieval and we applied our approach to derive new globally convergent schemes for this class.

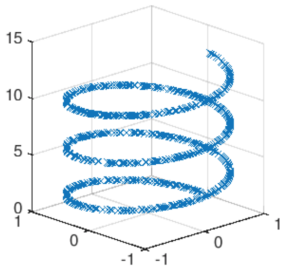
Low-rank reconstruction [9] This work is in the line of the previous one. The problem consists of solving inverse problems or minimizing losses with rank constraints, which is a form of algebraic sparsity. Using a model of Burer-Monteiro, we provided a universal kernel for treating such problems with large steps quartic methods. We applied our findings to symmetric nonnegative matrix factorization which is a key approach to probabilistic clustering or graph clustering. We also considered Euclidean distance matrix completion which is a fundamental problem with applications in sensor network localization and the study of the conformation of molecules. In that case, we developed new specific geometries (Gram Kernels) which provide remarkably fast results. Additionally, we proposed a dynamical update strategy, called DynNoLips, that allows to increase the step size beyond the conservative value predicted by theory and thus take advantage of local regularity of the function.

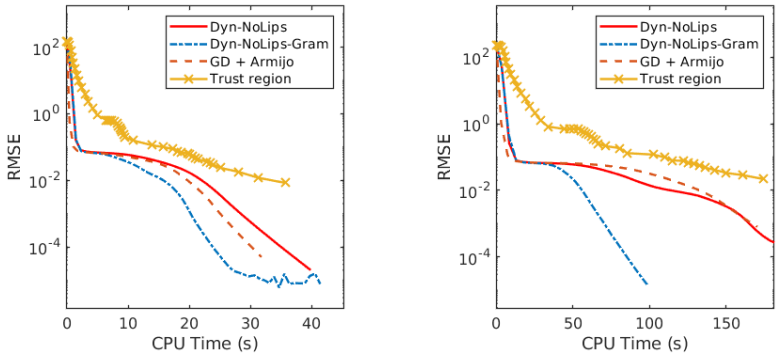
Let us provide an example that illustrates our findings on NoLips algorithm. We wish to recover the position of n points $X_1^*, \dots, X_n^*$ in $\mathbb{R}^r$ from an incomplete set of pairwise distances $\left\{d_{ij} = \|X_i^* - X_j^*\|^2 \,|\, (i,j) \in \Omega\right\}$. We then face the following *quadratic* problem:

$$\min_{X \in \mathbb{R}^{n \times r}} f(X) = \sum_{(i,j) \in \Omega} \left(\|X_i - X_j\|^2 - d_{ij}\right)^2$$

for which we used a Gram kernel as detailed in [9].

Experiments on synthetic `Helix` dataset with 10% known distances, dimension $r = 3$.





(a) $n = 2000, r = 3$ (b) $n = 5000, r = 3$

Figure 2: Euclidean matrix completion problems on the `Helix` dataset, with 10% known distances and two different problem sizes. We present the normalized RMSE over the full distance matrix versus CPU time. The results are averaged over 10 random initializations.

Lower bound for complexities [11] We obtained surprising negative results, a generic NoLips method cannot reach a precision ϵ with less than $O(1/\epsilon)$ iterations. This contrasts with what was known in the Euclidean case where $O(1/\sqrt{\epsilon})$ is a sharp bound. Our proof relies on a tricky counter-example which was guessed by computer-aided method. The counter-example features very awkward level sets so that the progress at each step is very limited whatever the descent strategy. On the bright side this also shows that the complexity provided in our paper [A] on NoLips is optimal (up to a constant).

2.2 Negative results [4]

In a recent article accepted at Math. Prog., we provided counterexamples to some old-standing optimization problems in the smooth convex coercive setting. These examples

are based on general smooth convex interpolation results. Given a decreasing sequence of positively curved C^p convex compact sets in the plane, we provided a level set interpolation of a C^p smooth convex function where $p \geq 2$ is arbitrary. If the intersection is reduced to one point our interpolant has a positive definite Hessian, otherwise it is positive definite out of the solution set.

Exact line search: the case of smooth convex coercive functions Exact line search constitutes a natural way to use gradient descent with large steps. In a smooth convex optimization context, this method is know to converge in value and has an $O(1/k)$ complexity. Our approach allows to produce a smooth convex function and a well defined exact line search sequence which does not converge.

Non convergence of NoLips Using the interpolation technique we developed we produced a continuous Legendre function h on a square S in the plane and a function f such that the algorithm (BG) produce a sequence x_k that does not converge. Note however that $f(x_k)$ converges to $\min_S f$ in $O(1/k)$.

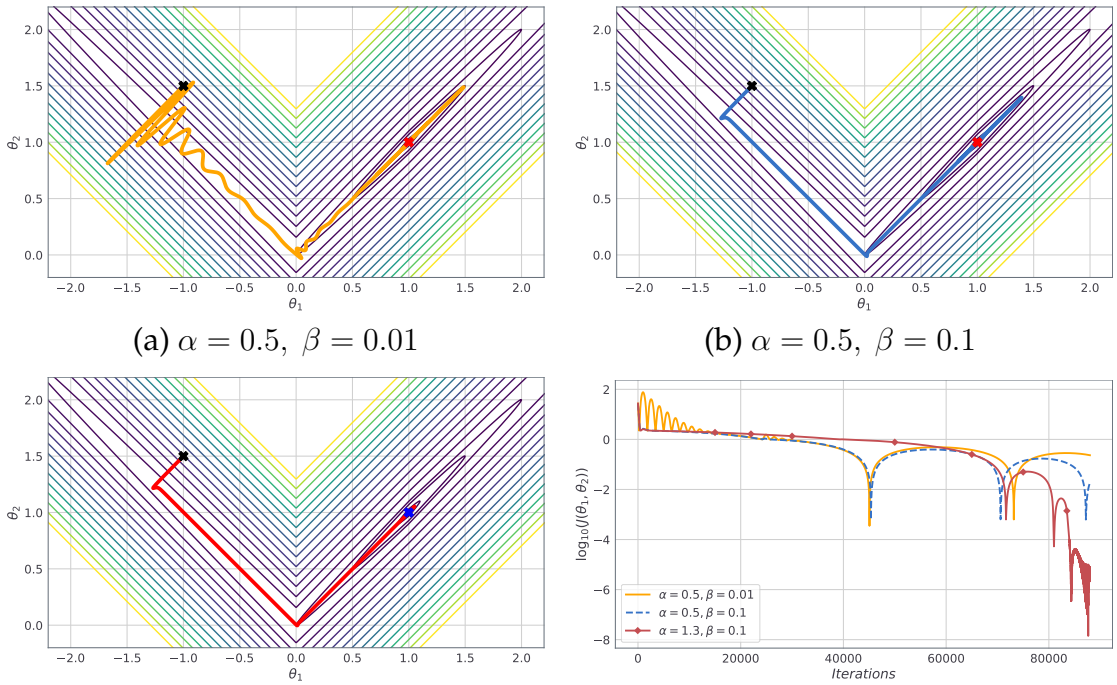
The two counterexamples above provides a fresh insight into old problems and raises the question of finding new rigidity assumptions to avoid these surprising pathologies.

2.3 Stochastic algorithms

Longer steps in Deep Learning [7] In Deep Learning the classical SGD method is applied using steps in $O(1/\sqrt{k})$ where k is the epoch counter. Using the theory of Benaim-Hofbauer-Sorin we showed that this bound is pessimistic and much more aggressive steps can be taken up to $o(1/\log(k))$. In this spirit, we also provided an algorithm stabilizing oscillations, called INNA, who is competing with the fastest methods for Deep Learning as ADAM. Our approach is inspired by the following continuous-time dynamical system introduced in a paper by Alvarez, Attouch, Bolte, Redont in 2003:

$$\underbrace{\ddot{\theta}(t)}_{\text{Inertial term}} + \underbrace{\alpha\,\dot{\theta}(t)}_{\text{Friction term}} + \underbrace{\beta\,\nabla^2 J(\theta(t))\dot{\theta}(t)}_{\text{Newtonian effects}} + \underbrace{\nabla J(\theta(t))}_{\text{Gravity effect}} = 0, \quad \text{for } t \in [0, +\infty),$$

where t is the time parameter which acts as a continuous epoch counter, J is a given loss function (e.g., the empirical loss in DL applications), for now assumed C^2 (twice-differentiable), with gradient ∇J and Hessian $\nabla^2 J$. We managed to adapt this method to nonsmooth Deep Learning problem by circumventing smoothness issues through a change of variable allowing as well “mini-batching". Figure 3 gives an idea of the local geometric intelligence of our method.

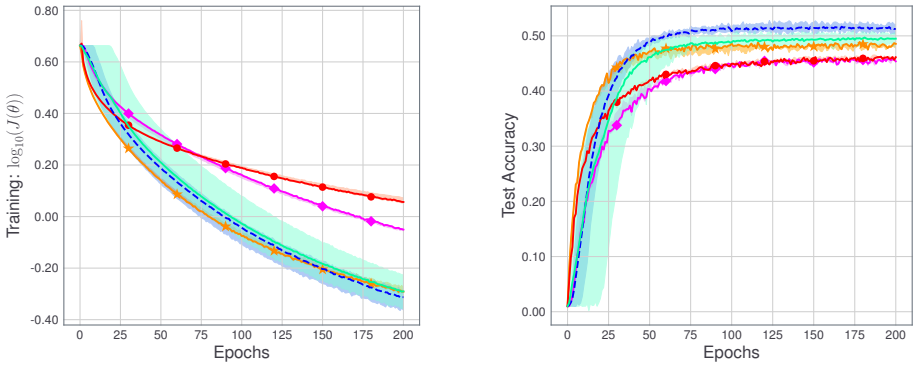


(a) $\alpha = 0.5,\ \beta = 0.01$ (b) $\alpha = 0.5,\ \beta = 0.1$

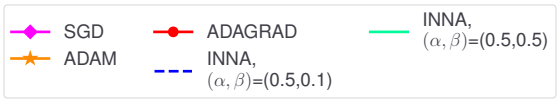
(c) $\alpha = 1.3,\ \beta = 0.1$ (d) Objective function in log-scale

Figure 3: Illustration of the role of the hyper-parameters of INNA on the non-smooth function $J(\theta_1, \theta_2) = 100(\theta_2 - |\theta_1|)^2 + |1 - \theta_1|$. The results are simulated using a full-batch version of the algorithm for three choices of hyper-parameters α and β. Subplot (d) displays the values of the objective function for the three settings considered.

Below we present some results on image classification: Classification of 60000 images in 100 categories with a moderately large neural network called NiN.



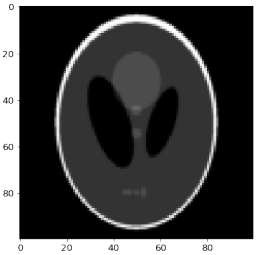
(a) $n = 2000, r = 3$ (b) $n = 5000, r = 3$



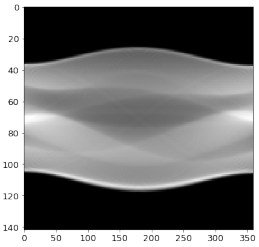
Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction [10] We study stochastic Bregman gradient methods for minimizing functions satisfying the relative regularity property. First, we show that the standard stochastic Bregman gradient method converges to a neighborhood of the optimum and oscillates because of the noise in the stochastic gradient estimate. A usual strategy to counter the effect of this noise is to use decreasing step size values, which causes the method to converge slowly. We rather propose to apply variance reduction techniques in order to use a fixed step size and obtain a fast convergence rate, at the expense of needing more memory to store previously computed gradients.

Let us provide an illustration of our algorithms BSGD and BSAGA (algorithm MU below is the state of the art method) for inverse problem with Poisson noise

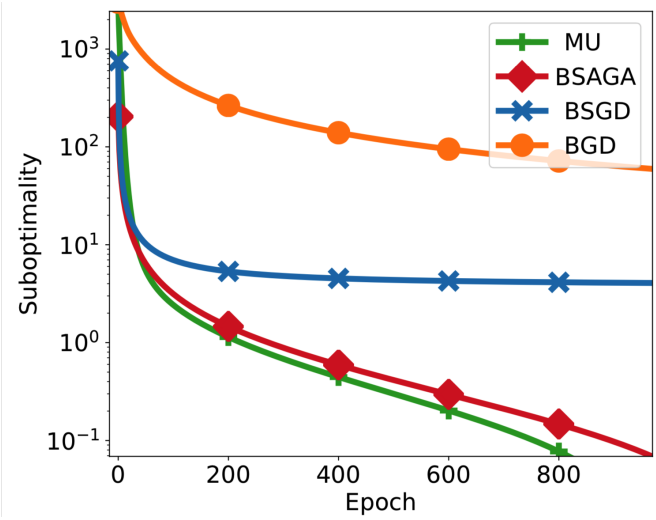
$$f(x) = D_{KL}(b, Ax), \quad h(x) = \sum_{i=1}^{d} -\log x^i.$$



Original signal x^*



Sinogram Ax^*



3 Aggressive steps in mini batch minimization [14]

We consider here nonconvex finite sum optimization, which is a typical problem arising in Deep Learning . Modern methods for this type of problems use minibatches and analyze the convergence through stochastic lenses.

A typical step size scheme for stochastic with replacement sampling decays like $1/\sqrt{k}$ where k is the iteration counter. In the context of "without replacement sampling (incremental methods)", we show that a less agressive step size strategy allows to obtain a faster convergence rate. Furthermore, this can be implemented in way which is adaptive to smoothness constants of the problem.

Adagrad step size has became a widespread preconditioning tool in machine learning. It adapts the steps sizes in a coordinatewise fashion, leaving the possibility to perform large steps for certain blocks of variables while keeping small steps for others (which

was also the spirit of our production [2]). We prove that the sequences generated by this algorithm converge in a smooth convex optimization context.

4 Conclusion and future works

The subject and the research presented in this proposal had a substantial impact on the optimization community since it gathers already more than 250 Google Scholar citations at this day (including our initial publication [A]).

The identification of an obstruction to acceleration for general kernels suggests focusing on more specific problems with concrete kernels: Boltzmann entropy, power type functions, or Burg's entropy. The acceleration problem for specific kernels has also attracted a lot of interest and still does. We are currently investigating this issue.

We also investigated a new research line related to AI. We are indeed working to design large steps in Deep Learning training and even in GANs (see e.g., [3]). This is a delicate matter because steps must vanish when activation functions are nonsmooth (eg ReLU, maxpool). On the other hand, qualification conditions are generally absent so that automatic differentiation does not always provide subgradient.

References

The founding paper is:

[A] H.H. Bauschke, J. Bolte, M. Teboulle, *A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications*, Mathematics of Operations Research, 42 (2), pp. 330–348, 2017

[1] H.H. Bauschke, J. Bolte, J. Chen, M. Teboulle, X. Wang, *On Linear Convergence of Non-Euclidean Gradient Methods without Strong Convexity and Lipschitz Gradient Continuity,* Journal of Optimization Theory and Applications, 2019

[2] J. Bolte, Z. Chen, E. Pauwels, *The multiproximal linearization method for convex composite problems*, Mathematical Programming, 2019.

[3] J. Bolte, L. Glaudin, E. Pauwels, M. Serrurier. *A Holderian backtracking method for min-max and min-min problems*. Submitted 2020.

[4] J. Bolte, E. Pauwels, *Curiosities and counterexamples in smooth convex optimization*, accepted in Math. Prog.

[5] J. Bolte, E. Pauwels, R. Rios-Zertuche, Long term dynamics of the subgradient method for Lipschitz path differentiable functions. Preprint

[6] J. Bolte, S. Sabach, M. Teboulle, Y. Vaisbourd, *First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems*, SIAM Journal on Optimization 28(3), pp. 2131–2151, 2018.

[7] C. Castera, J. Bolte, C. Févotte, E. Pauwels, *An Inertial Newton Algorithm for Deep Learning*, Journal of Machine Learning Research

[8] C. Castera, J. Bolte, C. Févotte, E.Pauwels. *Second-order step-size tuning of SGD for non-convex optimization*, Submitted 2021.

[9] RA. Dragomir, A. d'Aspremont, J. Bolte, *Quartic First-Order Methods for Low-Rank Minimization*, Journal of Optimization Theory and Applications 189(2), pp. 341–363, 2021.

[10] RA. Dragomir, M. Even, H. Hendrikx, *Fast Stochastic Bregman Gradient Methods: Sharp Analysis and Variance Reduction*, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, pp. 2815–2825, 2021.

[11] R. Dragomir, A. Taylor, A. d'Aspremont, J. Bolte, *Optimal Complexity and Certification of Bregman First-Order Methods*, Math. Prog., 2021.

[12] E. Pauwels, *Incremental without replacement sampling in nonconvex optimization*. Journal of Optimization Theory and Applications 190(1) pp 274–299, 2021.

[13] R. Rios-Zertuche, *Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions*. Preprint.

[14] C. Traoré, E. Pauwels, *Sequential convergence of AdaGrad algorithm for smooth convex optimization*, Operations Research Letters 49 (4), 452-458 2021.