ISE 403 Project

Nichole Tsz-Ching Chow

December 8, 2025

Abstract

Subgradient method is widely used for solving nonsmooth optimization problems. Its appliction widely spans various domains such as machine learning, signal processing, and operations research. In this paper, we review some recent advances in the convergence analysis of the subgradient method, focusing on its application to a wider range of nonsmooth optimization problems. We discuss various variants of the subgradient method and their convergence properties, highlighting key developments in the field. The paper aims to provide insights into the current state of research on the subgradient method and its potential applications in various domains.

1 Introduction

The purpose of this paper is to introduce the subgradient method for nonsmooth optimization and go through some recent advances in it tackling wider range of nonsmooth optimziation problem and its convergence analysis. Recently, some break through as been made in subgradient methods in tracking weakly convex [4], non-Lipschitz [6] objective functions. Instead of conducting original research, we will provide a comprehensive review of the existing literature on the subgradient method, focusing on its convergence analysis.

Subgradient method is one of the most fundamental algorithms for solving nonsmooth optimization problems. Subgradient method is first proposed by Shor et al. [16, 17] in the 1960s. Instead of using the gradient of the objective function, which may not exist at certain points of nonsmooth functions, it utilizes subgradients, which are generalized gradients that can be defined for nonsmooth functions. Similar to the gradient descent method for smooth optimization, the subgradient method iteratively updates the solution by moving in the direction of a subgradient, scaled by a step size. There are various applications of subgradient methods in machine learning [15, 9, 14], signal processing [18], and operations research [10].

Despite its wide applicability, extending the scope of appliction of subgradient methods to more general nonsmooth optimization problems and improving its convergence has been a subject of extensive research. Unlike the gradient descent method which is a descent method, the subgradient method does not

guarantee a monotonic decrease in the objective function value at each iteration [12]. This makes the convergence analysis more challenging. Over the years, researchers have proposed various step size strategies and modifications to improve the convergence rate of the subgradient method. Recent advances in this area have led to a better understanding of the algorithm's behavior and its performance under different conditions.

In this paper, we will discuss different variants of the subgradient method and their convergence properties, which tackle wider range of nonsmooth optimization problems. The rest of the paper is organized as follows. In Section 2, we introduce the notations and preliminaries used in this paper. In Section 3, we review some recent advances in the convergence analysis of the subgradient method.

2 Notations and Preliminaries

2.1 Notations

Throughout the paper, we represent scalars, vectors, and matrices using lowercase letters, bold lowercase letters, and uppercase letters, respectively. We use $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}^n$, and $\mathbb{R}^{m\times n}$ to represent the set of real numbers, nonegative real numbers, n-dimensional real vectors, and $m \times n$ real matrices, respectively. For a real matrix $M \in \mathbb{R}^{m\times n}$, we denote $M^\top$ as the transpose of M. When $m = n$, the minimal and maximal eigenvalues of M are represented as $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$, respectively. For a symmetric matrix M, $M \succeq 0$ means M is positive semidefinite (PSD).

We use $\langle\cdot,\cdot\rangle$, and $\|\cdot\|$ to denote the inner product and norm induced from the inner product. For an extended real-valued function f, the domain of f is defined as $\mathrm{dom}(f) := \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) < \infty\}$. A function f is proper if $\mathrm{dom} f \neq \emptyset$ and $f(\mathbf{x}) > -\infty$ for all $x \in \mathrm{dom}(f)$, and is closed if it is lower semicontinuous, which is when $\mathrm{epi}(f)$ is closed, where $\mathrm{epi}(f) := \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} \mid f(\mathbf{x}) \leq t\}$ is the epigraph of f. For any subset $S \subseteq \mathbb{R}^n$ and any point $\mathbf{x} \in \mathbb{R}^n$, the distance from $\mathbf{x}$ to S is defined by $\mathrm{dist}(\mathbf{x}, S) := \inf\big\{\|\mathbf{y} - \mathbf{x}\| \;\big|\; \mathbf{y} \in S\big\}$, and $\mathrm{dist}(\mathbf{x}, S) = \infty$ when $S = \emptyset$.

2.2 Preliminaries of Nonsmooth Optimization

In this section, we will go through some fundamental knowledge on nonsmooth optimization. For a nonsmooth optimization problem, it can be strongly convex, convex, or nonconvex.

Definition 1. (**Strong convexity**) *Let $\lambda > 0$, a convex function f is said to be λ-strongly convex over $\mathcal{X}$ if and only if $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2$ for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{X}_C$.*

Then, we will go through the key concept of subdifferentials, which is a generalized idea of gradient when the function $f \notin \mathcal{C}^1$.

Definition 2. (**Subdifferentials**) [1, 2] *For a proper and lower semicontinuous function* $f : \mathbb{R}^n \to (-\infty, \infty]$,

(i) given $\mathbf{x} \in \mathrm{dom}(f)$*, the Fréchet subdifferential of* f *at* $\mathbf{x}$*, expressed as* $\widehat{\partial} f(\mathbf{x})$*, is the set of all vectors* $\mathbf{u} \in \mathbb{R}^n$ *satisfying*

$$\liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \to \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0, \tag{1}$$

and we set $\widehat{\partial} f(\mathbf{x}) = \emptyset$ *when* $\mathbf{x} \notin \mathrm{dom}(f)$.

(ii) (limiting-)subdifferential of f *at* $\mathbf{x}$*, written by* $\partial f(\mathbf{x})$*, is defined by*

$$\partial f(\mathbf{x}) := \{\mathbf{u} \in \mathbb{R}^n \mid \exists\ \mathbf{x}^k \to \mathbf{x},\ \text{s.t. } f(\mathbf{x}^k) \to f(\mathbf{x}) \text{ and } \widehat{\partial} f(\mathbf{x}^k) \ni \mathbf{u}^k \to \mathbf{u}\}. \tag{2}$$

(iii) a point $\mathbf{x}^*$ *is called (limiting-)critical point or stationary point of* f *if it satisfies* $0 \in \partial f(\mathbf{x}^*)$*, and the set of critical points of* f *is denoted by* $\mathrm{crit} f$.

Note that (1) implies that the property $\widehat{\partial} f(\mathbf{x}) \subseteq \partial f(\mathbf{x})$ immediately holds, and $\widehat{\partial} f(\mathbf{x})$ is closed and convex, meanwhile $\partial f(\mathbf{x})$ is closed [13, Theorem 8.6]. Also, the subdifferential (2) reduces to the gradient of f denoted by ∇f if f is continuously differentiable. Moreover, as mentioned in [13], if g is a continuously differentiable function, it holds that $\partial(f + g) = \partial f + \nabla g$.

In Section 3, we will also cover the projected algorithm, for which Euclidean projection is defined as follows

Definition 3. (Euclidean projection) *Let* $\mathcal{X} \subseteq \mathbb{R}^n$ *be a nonempty, closed, convex set. Then, the Euclidean projection of a point* $\mathbf{x}$ *onto* $\mathcal{X}$ *is defined as*

$$P_X(\mathbf{y}) = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y} - \mathbf{x}\|$$

3 Literature Review

Subgradient Method. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a convex set and f be a convex function with $\mathrm{dom}(f) \subseteq \mathcal{X}$. For the minimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}), \tag{3}$$

gradient descent is commonly applied when f is smooth, meaning that its gradient is L-Lipschitz; that is,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \quad \forall (\mathbf{x}, \mathbf{y}) \in \mathrm{dom}(f) \times \mathrm{dom}(f). \tag{4}$$

When the objective function f is nonsmooth, the gradient does not exist, and hence the subgradient of f is computed instead of the gradient. The set of subgradients of the function f at a point $\mathbf{x}$, denoted as $\partial f(\mathbf{x})$ is defined as

$$\partial f(\mathbf{x}) = \{\mathbf{g} \in \mathbb{R}^n \text{ such that } f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle\}.$$

For solving the minimization problem (3), the subgradient method was first proposed by Shor [17] in the 1960s. The iterative scheme of the subgradient method first computes a subgradient $\mathbf{g} \in \partial f(\mathbf{x}^k)$ and updates the next iterate $\mathbf{x}^{k+1}$ as follows

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \mathbf{g}^k,$$

where α_k is the stepsize for iteration k. It is commonly known that, the lower bound [11, Theorem 3.2.1] on the subgradient method with a B-Lipschitz continuous function f is

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) \leq \frac{BR}{2(2+\sqrt{k+1})},$$

where $\mathbf{x}^*$ is the optimal solution and $\|\mathbf{x}^0 - \mathbf{x}^*\| \leq R$ for some $R > 0$.

The subgradient method remains an active topic of research, as nonsmoothness is prevalent in modern machine learning tasks such as regression, classification, and matrix completion.

Kelley's Cutting-plane Method. Recently, Drori and Teboulle [5] were motivated by Kelley's cutting plane method [8, 3], which keeps a polyhedral model of the objective and updates it at each iteration using first-order information at the point predicted to minimize the objective. Their approach established a convergence rate of

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) \leq \frac{BR}{\sqrt{k+1}}.$$

Non-traditional Stepsize Scheme Then, Jain et al. [7] have further improved the convergence rate to $\mathcal{O}(\frac{1}{\sqrt{k}})$ with a non-standard step size scheme. They have proposed a modified step size scheme $(\alpha_t)_{t=1}^T$ as

$$\alpha^t := 2^{-i}\gamma^t, \quad \forall T_i < t \leq T_{i+1}, \quad 0 \leq i \leq k,$$

where $(\gamma_t)_{t=1}^T$ is the subgradient descent stepsize sequence. This ensures the step sizes eventually decay fast enough for the iterates to approach the optimum $\mathbf{x}^*$. However, their result holds only when the feasible set $\mathcal{X}$ is bounded. Their proposed method has a convergence rate of

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) \leq \frac{15BD}{\sqrt{k+1}},$$

where $D = \max_{(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$. This is desirable compared to the methods mentioned above, as it provides an upper bound on the convergence rate of the subgradient method and establishes the rate even in the worst-case scenario.

Exact Convergence. Later, Zamani et al. [19] have relaxed the limitation of compactness of domains $\mathcal{X}$ and derived a upper bound on the convergence rate of $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \frac{BR}{\sqrt{k+1}}$ for the last iterate of the subgradient method with an easily satisfied requirement on step sizes. This has significantly improved the convergence in Jain et al. [7]. Their convergence analysis relies on the

construction of a reference point $\mathbf{z}^k$ to have the quality $f(\mathbf{x}^k)-f(\mathbf{z}^{k-1})$ bounding the optimality gap $f(\mathbf{x}^k) - f(\mathbf{x}^*)$. After k iterations, with their proposed step size, the last-iterate accuracy will be smaller than

$$\frac{BR}{\sqrt{B+1}}\sqrt{1+\frac{\log(k)}{4}}.$$

4 Biography

Nichole Tsz-Ching Chow is a first-year Ph.D. student in the Department of Industrial and Systems Engineering at Lehigh University, focusing on nonconvex nonsmooth optimization with applications in machine learning. She has been awarded a Rossin/Parker Fellowship for 2025. She received her B.Sc. in Mathematics from the Chinese University of Hong Kong (CUHK) in 2023. During her undergraduate studies, she was awarded the prestigious Professor Charles K. Kao Research Exchange Scholarship to work as a research intern at the University of Tennessee under the supervision of Dr. Kwai Wong on large-scale parallel computing. She later completed an M.Phil. in Mathematics from CUHK in 2025, advised by Prof. Tieyong Zeng.

References

[1] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[2] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

[3] Elliot Ward Cheney and Allen A Goldstein. Newton's method for convex programming and tchebycheff approximation. *Numerische Mathematik*, 1(1):253–268, 1959.

[4] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[5] Yoel Drori and Marc Teboulle. An optimal variant of kelley's cutting-plane method. *Mathematical Programming*, 160(1):321–351, 2016.

[6] Benjamin Grimmer and Danlin Li. Some primal-dual theory for subgradient methods for strongly convex optimization. *Mathematical Programming*, pages 1–30, 2025.

[7] Prateek Jain, Dheeraj M Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021.

[8] James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[9] Jyrki Kivinen, Alex J Smola, and Robert C Williamson. Online learning with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1083–1101, 2004.

[10] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[11] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[12] Boris T Polyak. Subgradient methods: a survey of soviet research. *Nonsmooth optimization*, pages 5–29, 1977.

[13] R Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

[14] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[15] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, 2007.

[16] Naum Z Shor. An application of the method of gradient descent to the solution of the network transportation problem. *Materialy Naucnovo Seminara po Teoret i Priklad. Voprosam Kibernet. i Issted. Operacii, Nucnyi Sov. po Kibernet, Akad. Nauk Ukrain. SSSR, vyp*, 1:9–17, 1962.

[17] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

[18] Isao Yamada and Nobuhiko Ogura. Adaptive projected subgradient method and its applications to set theoretic adaptive filtering. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 1, pages 600–606. IEEE, 2003.

[19] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025.