OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Incremental Methods for Weakly Convex Optimization

Xiao Li LIXIAO@CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen

Zhihui Zhu ZHIHUI.ZHU@DU.EDU
University of Denver

Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

Jason D. Lee JASONLEE@PRINCETON.EDU
Princeton University

Abstract

We consider incremental algorithms for solving *weakly convex* optimization problems, a wide class of (possibly nonsmooth) nonconvex optimization problems. We will analyze incremental (sub)-gradient, proximal point, and prox-linear methods. We show that the convergence rate of the three incremental algorithms is $\mathcal{O}(k^{-1/4})$ under weakly convex setting. This extends the convergence theory of incremental methods form convex optimization to nonsmooth nonconvex regime. When the weakly convex function satisfies an additional regularity condition called *sharpness* property, we show that all the three incremental algorithms with a geometrical diminishing stepsize rule and an appropriate initialization converge even *linearly* to the optimal solution set. We conduct experiments on robust matrix sensing and robust phase retrieval to illustrate the superior convergence performance of incremental methods.

1. Introduction

We consider incremental methods for addressing the finite-sum optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{x}), \tag{1}$$

where each component function f_i is *weakly convex*. Recall that a function σ is said to be weakly convex if $\sigma(\cdot) + \frac{\tau}{2}\|\cdot\|^2$ is convex for some constant $\tau \geq 0$ [19]. We assume the global minimum set $\mathcal{X}$ of (1) is nonempty and denote $f^\star$ as its minimal function value. It is worthy to mention that f in (1) can be *nonsmooth* and *nonconvex* under the weakly convex setting, covering rich applications in practice. As an illustration, let us present two motivating applications, in which nonsmooth formulations have clear advantages over smooth ones.

Robust Matrix Sensing [13] One fundamental computational task in machine learning and signal processing is to recover a Positive Semi-Definite (PSD) low-rank matrix $\boldsymbol{X}^\star \in \mathbb{R}^{n\times n}$ with $\mathrm{rank}(\boldsymbol{X}^\star) = r \leq n$ from a small number of corrupted linear measurements

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^\star) + \boldsymbol{s}^\star, \tag{2}$$

where $\mathcal{A}$ is a linear measurement operator consisting of a set of sensing matrices $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_m$ and $\boldsymbol{s}^\star$ is a sparse outliers vector. An effective approach to recover the low-rank matrix $\boldsymbol{X}^\star$ is by using a factored representation of the matrix variable [3] (i.e., $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ with $\boldsymbol{U} \in \mathbb{R}^{n\times r}$) and employing an ℓ_1-loss function to robustify the solution against outliers:

$$\underset{\boldsymbol{U}\in\mathbb{R}^{n\times r}}{\text{minimize}} \ \frac{1}{m}\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}})\|_1 = \frac{1}{m}\sum_{i=1}^{m}|y_i - \langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\rangle|. \tag{3}$$

Direct calculation shows that each component function in (3) is weakly convex.

Robust Phase Retrieval [8, 9] Robust phase retrieval aims to recover a signal $\boldsymbol{x}^\star \in \mathbb{R}^n$ from its corrupted magnitude-wise measurements

$$\boldsymbol{b} = |\boldsymbol{A}\boldsymbol{x}^\star|^2 + \boldsymbol{s}^\star, \tag{4}$$

the operator $|.|^2$ in (4) means taking modulus and then squaring coordinate-wise. Here, the matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the measurement matrix and $\boldsymbol{s}^\star \in \mathbb{R}^m$ is the sparse outliers vector. The work [8] formulates the following problem for recovering both the sign and magnitude information of $\boldsymbol{x}^\star$:

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{m}\left\||\boldsymbol{A}\boldsymbol{x}|^2 - \boldsymbol{b}\right\|_1 = \frac{1}{m}\sum_{i=1}^{m}\left||\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle|^2 - b_i\right|. \tag{5}$$

It is straightforward to verify that each component function in (5) is weakly convex.

1.1. The Algorithms

Incremental methods play an important role in large-scale optimization problems such as the training of deep neural networks. In this paper, we study a family of incremental methods, including incremental (sub)-gradient, proximal point, and prox-linear methods. In each time, incremental methods update the iterate with only *one component function* f_i selected according to a cyclic order—i.e., select component function *sequentially* from f_1 to f_m and repeating such process *cyclically*. To be more specific, in $k+1$-th iteration, incremental algorithms start with $\boldsymbol{x}_{k,0} = \boldsymbol{x}_k$, and then update $\boldsymbol{x}_{k,i}$ using f_i using certain method for all $i = 1, \cdots, m$, giving $\boldsymbol{x}_{k+1} = \boldsymbol{x}_{k,m}$. The following three incremental algorithms differ from each other in the update of $\boldsymbol{x}_{k,i}$.

Incremental (sub)-gradient method:

$$\boldsymbol{x}_{k,i} = \boldsymbol{x}_{k,i-1} - \mu_k \widetilde{\nabla} f_i(\boldsymbol{x}_{k,i-1}), \tag{6}$$

where $\widetilde{\nabla} f_i$ is any (sub)-gradient belongs to the Fréchet subdifferential ∂f_i (see (11)).

Incremental proximal point method:

$$\boldsymbol{x}_{k,i} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{argmin}} \ f_i(\boldsymbol{x}) + \frac{1}{2\mu_k}\|\boldsymbol{x} - \boldsymbol{x}_{k,i-1}\|_2^2. \tag{7}$$

Incremental prox-linear method: We now consider a special class of weakly convex functions that are of the composite form (cf. (3) and (5))

$$f(\boldsymbol{x}) = h(c(\boldsymbol{x})) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} h_i(c_i(\boldsymbol{x})), \tag{8}$$

where each $h_i : \mathbb{R}^d \to \mathbb{R}$ is a (possibly nonsmooth) Lipschitz convex mapping and $c_i : \mathbb{R}^n \to \mathbb{R}^d$ is a smooth function with Lipschitz continuous Jacobian. We denote by

$$f_i(\boldsymbol{x}; \boldsymbol{x}_{k,i-1}) = h_i \left(c_i(\boldsymbol{x}_{k,i-1}) + \nabla c_i(\boldsymbol{x}_{k,i-1})^T (\boldsymbol{x} - \boldsymbol{x}_{k,i-1}) \right). \tag{9}$$

The incremental prox-linear method update as

$$\boldsymbol{x}_{k,i} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^n} \; f_i(\boldsymbol{x}; \boldsymbol{x}_{k,i-1}) + \frac{1}{2\mu_k} \|\boldsymbol{x} - \boldsymbol{x}_{k,i-1}\|_2^2. \tag{10}$$

1.2. Prior Arts

Though incremental methods are broadly used, its theoretical insights are far from being well understood. The main prior achievements for analyzing incremental methods are based on convexity assumption. The starting work may date back to [20] for solving least square problems. Then various works [2, 10, 11, 14, 18] extended incremental gradient descent to training shallow linear neural networks and solving other smooth convex problems. When the component function f_i is convex but nonsmooth, incremental (sub)-gradient method was proposed in [12] for solving finite-sum nonsmooth convex optimization problems. Later, Nedic and Bertsekas [15, 16] provided convergence results for incremental (sub)-gradient method using different stepsize rules.

On the other hand, *stochastic methods* for weakly convex optimization are recently studied in [5, 6]. We remark that incremental methods are fundamentally different from the stochastic ones, as the former are essentially *deterministic* algorithms.

To the best of our knowledge, there is no convergence result for incremental methods when the function f in (1) is nonsmooth and nonconvex. Thus, it is fundamentally important to ask:

> *Are the incremental methods studied in this paper guaranteed to converge if f in (1) is nonsmooth and nonconvex? If yes, what is the convergence rate?*

In this paper, we answer this question positively under the assumption that each component function f_i in (1) is weakly convex. Our work builds upon the original proofs in [15, 16] which analyzed the convergence of incremental (sub)-gradient method when used to solve nonsmooth *convex* problems. We also adapt the surrogate stationarity measure from [5] for analyzing weakly convex minimization.

2. Main Convergence Results

2.1. Preliminaries

Subdifferential Since f can be nonsmooth, we utilize tools from generalized differentiation. The (Fréchet) subdifferential of a function σ at $\boldsymbol{x}$ is defined as (see, e.g., [17])

$$\partial \sigma(\boldsymbol{x}) := \left\{ \widetilde{\nabla} \sigma(\boldsymbol{x}) \in \mathbb{R}^n : \liminf_{\boldsymbol{y} \to \boldsymbol{x}} \frac{\sigma(\boldsymbol{y}) - \sigma(\boldsymbol{x}) - \langle \widetilde{\nabla} \sigma(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|} \geq 0 \right\}, \tag{11}$$

where each $\widetilde{\nabla} \sigma(\boldsymbol{x}) \in \partial \sigma(\boldsymbol{x})$ is called a subgradient of σ at $\boldsymbol{x}$.

Sharpness Property The sharpness property characterizes how fast the function increases when x is away from the set of global minima. We say that a mapping $\sigma : \mathbb{R}^n \to \mathbb{R}$ is α-sharp where $\alpha > 0$ if (see, e.g., [4])

$$\sigma(\boldsymbol{x}) - \sigma^\star \geq \alpha \operatorname{dist}(\boldsymbol{x}, \mathcal{X}) \tag{12}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$. Here $\mathcal{X}$ denotes the set of global minimizers of σ, $\sigma^\star$ represents the minimal value of σ, and $\operatorname{dist}(\boldsymbol{x}, \mathcal{X})$ is the distance of $\boldsymbol{x}$ to $\mathcal{X}$, i.e., $\operatorname{dist}(\boldsymbol{x}, \mathcal{X}) = \inf_{\boldsymbol{x}' \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$.

Moreau Envelope We will utilize the concept of Moreau envelope for defining stationarity measure. For any $\lambda > 0$, the Moreau envelope of $\sigma : \mathbb{R}^n \to \mathbb{R}$ is defined as (see [17, Definition 1.22])

$$\sigma_\lambda(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathbb{R}^n} \ \sigma(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{13}$$

The corresponding proximal mapping is defined as

$$\operatorname{prox}_{\lambda,\sigma}(\boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{y} \in \mathbb{R}^n} \ \sigma(\boldsymbol{y}) + \frac{1}{2\lambda}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{14}$$

2.2. Global Convergence

In this section, we study the iteration complexities of the incremental methods under weakly convex setting.

Assumptions and Stationarity Measure In addition to weak convexity, the following is assumed throughout this section, which is standard for analyzing incremental methods; see, e.g., [1, 15, 16].

Assumption 1 *(bounded subgradients) For any $i \in \{1, \cdots, m\}$, there exists a constant $L > 0$, such that $\|\widetilde{\nabla} f_i(\boldsymbol{x})\| \leq L$, for all $\widetilde{\nabla} f_i(\boldsymbol{x}) \in \partial f_i(\boldsymbol{x})$ and $\boldsymbol{x} \in \operatorname{dom} f$.*

Due to the nonsmoothness of the objective function, we borrow ideas from the recent works [5, 7] on weakly convex minimization, which propose to use the gradient of the Moreau envelope of the weakly convex function at hand as a surrogate stationarity measure. Formally, we have [5]

$$\begin{cases} \dfrac{1}{\lambda}\left\|\boldsymbol{x} - \overline{\boldsymbol{x}}\right\| = \left\|\nabla \sigma_\lambda(\boldsymbol{x})\right\|, \\ \operatorname{dist}\left(0, \partial\sigma(\overline{\boldsymbol{x}})\right) \leq \left\|\nabla \sigma_\lambda(\boldsymbol{x})\right\|, \end{cases} \tag{15}$$

where σ_λ and $\overline{\boldsymbol{x}} = \operatorname{prox}_{\lambda,\sigma}(\boldsymbol{x})$ are defined in (13) and (14), respectively. Clearly, we see from (15) that $\boldsymbol{x}$ is a stationary point of problem (1) when $\|\nabla \sigma_\lambda(\boldsymbol{x})\| = 0$. Thus, we call $\boldsymbol{x}$ an *ε-nearly stationary point* of problem (1) if $\|\nabla \sigma_\lambda(\boldsymbol{x})\| \leq \varepsilon$. The main result of this section is presented in the following theorem.

Theorem 1 *Suppose that Assumption 1 is valid. Let the constant $\widehat{\tau} > 2\tau$ and the sequence $\{\boldsymbol{x}_k\}$ be generated by any of the three incremental methods for solving (1) with arbitrary initialization. Suppose further the stepsize $\mu_k = \mu = \frac{1}{m\tau\sqrt{N+1}}$ is constant for all $k \geq 0$, where integer N is the total iteration number. Then we have*

$$\begin{aligned} \min_{0 \leq k \leq N} \left\|\nabla f_{1/\widehat{\tau}}(\boldsymbol{x}_k)\right\|^2 &\leq \frac{C_1}{\left(\frac{\widehat{\tau}}{2} - \tau\right)\sqrt{N+1}} + \frac{C_2}{\left(\frac{\widehat{\tau}}{2} - \tau\right)(N+1)} \\ &= \mathcal{O}\left(\frac{1}{\sqrt{N+1}}\right), \end{aligned} \tag{16}$$

where C_1 and C_2 are positive numerical constants.

In terms of iteration complexity, Theorem 1 implies that the incremental methods requires at most $\mathcal{O}(\varepsilon^{-4})$ number of iterations to obtain an ε-nearly stationary point.

2.3. Local Linear Convergence

In this section, we show that by exploiting the sharpness property of f, the incremental methods can even converge locally linearly for weakly convex minimization. In addition to weak convexity and Assumption 1, we make the following additional assumption in this section.

Assumption 2 *(sharpness) The function f in (1) is α-sharp; see Section 2.1 for definition.*

We now state the main result of this section in the following theorem, which indicates the local linear convergence rate of the incremental methods with a suitably designed geometrically decaying stepsize and an appropriate initialization.

Theorem 2 *Suppose that Assumption 1 and 2 are valid. Let the sequence of iterates $\{\boldsymbol{x}_k\}$ be generated by any of the three incremental methods with initialization $\boldsymbol{x}_0$ satisfying $\operatorname{dist}(\boldsymbol{x}_0, \mathcal{X}) \leq \frac{\alpha}{2\tau}$. If the stepsize μ_k in all the three incremental methods is updated as $\mu_k = \rho^k \mu_0$, with $0 < \mu_0 \leq \frac{\alpha^2}{5m\tau L^2}$, and $0 < \sqrt{1 - 2m\tau\mu_0 + \frac{5m^2\tau^2L^2}{\alpha^2}\mu_0^2} \leq \rho < 1$. Then, we have*

$$\operatorname{dist}(\boldsymbol{x}_k, \mathcal{X}) \leq \rho^k \cdot \frac{\alpha}{2\tau}, \quad \forall\, k \geq 0.$$

3. Simulations

In this section, we conduct a series of experiments on robust matrix sensing (2)-(3). We generate $\boldsymbol{U}^\star \in \mathbb{R}^{n\times r}$ and m sensing matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m \in \mathbb{R}^{n\times n}$ (which forms the linear operator $\mathcal{A}$) with *i.i.d.* standard Gaussian entries. The ground truth low-rank matrix is generated by $\boldsymbol{X}^\star = \boldsymbol{U}^\star \boldsymbol{U}^{\star\mathrm{T}}$. We generate the outliers vector $\boldsymbol{s}^\star \in \mathbb{R}^m$ by first randomly selecting pm locations, where p is the preset outliers ratio. Then, each of the selected location is filled with an *i.i.d.* mean 0 and variance 10 Gaussian entry, while the value on the remaining locations are set to 0. According to (2), the linear measurement $\boldsymbol{y} \in \mathbb{R}^m$ is generated by $y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\star \rangle + s_i^\star$, $i = 1, \ldots, m$. We set the parameters as $n = 50,\ r = 5,\ m = 5nr,\ p = 0.3$. In this case, it is shown in [13] that the global optimal solution set of (3) is exactly $\mathcal{U} = \{\boldsymbol{U}^\star \boldsymbol{R} : \boldsymbol{R} \in \mathbb{R}^{r\times r}, \boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}\}$. We compare our incremental (sub)-gradient method (ISGM) and incremental prox-linear method (IPL) with (sub)-gradient method (SGM), stochastic (sub)-gradient method (SGD), and stochastic prox-linear method (SPL). We tune the best stepsize decay factor β for each algorithm. The results are shown in Figure 1. One can observe that the incremental methods outperform other algorithms in terms of convergence speed.

References

[1] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163, 2011.

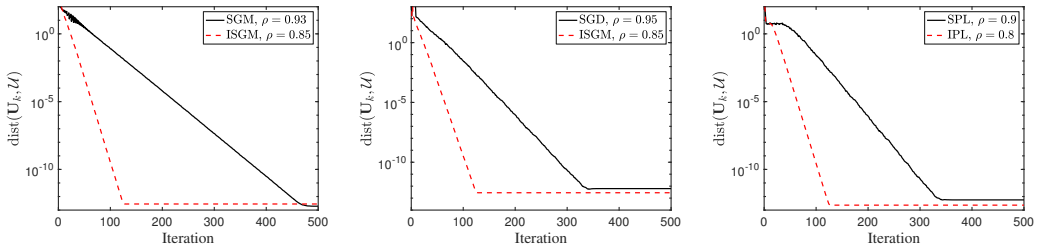

Figure 1: Performance on robust matrix sensing.

[2] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of the 34th IEEE Conference on Decision and Control*, volume 1, pages 560–564. IEEE Publ. Piscataway, NJ, 1995.

[3] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[4] James V Burke and Michael C Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

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

[6] Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

[7] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, pages 1–56, 2018.

[8] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*, 2017.

[9] Yonina C Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[10] Luigi Grippo. Convergent on-line algorithms for supervised learning in neural networks. *IEEE transactions on neural networks*, 11(6):1284–1299, 2000.

[11] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo Parrilo. Convergence rate of incremental gradient and newton methods. *arXiv preprint arXiv:1510.08562*, 2015.

[12] VM Kibardin. Decomposition into functions in the minimization problem. *Avtomatika i Telemekhanika*, (9):66–79, 1979.

[13] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Rene Vidal. Nonconvex robust low-rank matrix recovery. *arXiv preprint arXiv:1809.09237*, 2018.

[14] Zhi-Quan Luo. On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks. *Neural Computation*, 3(2):226–245, 1991.

[15] Angelia Nedić and Dimitri Bertsekas. Convergence rate of incremental subgradient algorithms. In *Stochastic optimization: algorithms and applications*, pages 223–264. Springer, 2001.

[16] Angelia Nedic and Dimitri P Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[17] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[18] Paul Tseng. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

[19] Jean-Philippe Vial. Strong and Weak Convexity of Sets and Functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

[20] Bernard Widrow and Marcian E Hoff. Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs, 1960.