Journal of Applied Mathematics and Physics, 2026, 14(4), 1612-1626
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Research on the Proximal Gradient Method for Composite Optimization Problems under Generalized Smoothness Assumptions

Lin Yang, Na Xian*

Key Laboratory of Optimization Theory and Applications, School of Mathematical Sciences, China West Normal University, Nanchong, China
Email: 1540260500@qq.com, *1564609301@qq.com

How to cite this paper: Yang, L. and Xian, N. (2026) Research on the Proximal Gradient Method for Composite Optimization Problems under Generalized Smoothness Assumptions. *Journal of Applied Mathematics and Physics*, **14**, 1612-1626.
https://doi.org/10.4236/jamp.2026.144076

Received: March 31, 2026
Accepted: April 20, 2026
Published: April 23, 2026

Abstract

The proximal gradient method (PGD) is an important approach for solving composite optimization problems consisting of the sum of a smooth function and a nonsmooth function. Classical convergence analysis of PGD typically assumes that the smooth function has a globally Lipschitz continuous gradient. In recent years, researchers have relaxed this assumption from various perspectives, thereby providing theoretical support for the application of PGD to more general problems. In particular, for unconstrained smooth optimization problems, Li *et al.* introduced the concept of $\ell(\cdot)$-smoothness, studied the convergence rates of classical gradient methods, and showed that under this generalized smoothness condition, the convergence rates of classical gradient methods remain consistent with those under the classical smoothness condition. Nevertheless, existing results are mostly focused on unconstrained smooth optimization, and the corresponding theoretical analysis of PGD for composite optimization problems still requires further development. To this end, this paper further investigates the convergence rates of PGD for solving composite optimization problems within the $\ell(\cdot)$-smoothness framework. Under the assumption that the smooth component in the composite optimization problem is convex, we prove that the sequence of function values generated by the constant-stepsize PGD under-smoothness achieves a convergence rate of $O(1/k)$.

Keywords

Generalized Smoothness, Proximal Gradient Method, Composite Optimization Problem

1. Introduction

This paper considers the following composite convex optimization problem:

$$\min_{x\in R^d} F(x) = f(x) + g(x) \tag{1}$$

where $f:\mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a proper closed convex function that is differentiable on the open effective domain $\mathcal{X} = \operatorname{dom} f$, $g:\mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a proper closed convex function satisfying $\operatorname{dom} g \subseteq \mathcal{X}$.

Problem (1) has wide applications in fields such as compressed sensing, sparse phase retrieval [1] [2], network quantization [3], and machine learning. Currently, researchers have proposed various classical methods for solving problem (1), including the subgradient algorithm, splitting algorithms [4], and the proximal gradient method [5].

The proximal gradient method (PGD) combines gradient descent with the proximal operator to solve optimization problems with nonsmooth regularization terms by decomposing complex objective functions. Its iterative scheme is as follows:

$$x^{t+1} = \operatorname{prox}_{\lambda_t, g}\left\{x^t - \nabla f(x)\right\},$$

where $\lambda_t > 0$ is the step size parameter, and $\operatorname{prox}_{\lambda_t, g}$ denotes the proximal operator of the function g defined as

$$\operatorname{prox}_{\lambda_t, g}(z) = \arg\min_{x\in\mathbb{R}^d}\left\{g(x) + \frac{1}{2\lambda_t}\|x - z\|^2\right\}.$$

In the convergence theory analysis of traditional PGD, it is often required that the gradient satisfy Lipschitz continuity. Under this smoothness condition, if f is convex, the algorithm can achieve a sublinear convergence rate of $O(1/k)$.

However, the classical smoothness condition of gradient Lipschitz continuity is relatively strict. For functions with rapidly changing curvature or uneven local geometric structures, this assumption is often difficult to satisfy. Particularly in machine learning and large-scale optimization, many objective functions, although differentiable or even twice differentiable, may not have globally Lipschitz continuous gradients. Even in some standard models, the global smoothness constant may be unbounded [6], such as the ℓ_2-regression function. Additionally, when the function f is an augmented Lagrangian function or the dual function of the original problem (which is not necessarily uniformly convex), the objective function may also exhibit nonsmooth or non-uniformly smooth characteristics [7] [8]. Therefore, how to establish algorithm convergence theory under conditions weaker than classical smoothness remains an important question.

Significant progress has been made in research on weakening the classical smoothness assumption. In 2020, Zhang *et al.* [9], based on an in-depth study of the training process of deep neural networks such as LSTM [10] and ResNet [11], first proposed a non-uniform smoothness condition—(L_0, L_1)-smoothness—to characterize the dependence between the variation of the function gradient and the gradient norm. Compared with the classical smoothness condition of gradient

Lipschitz continuity, this type of condition allows the local smoothness degree to vary with the position of the iteration point or the gradient norm, thus covering a broader class of functions, such as univariate polynomials and exponential functions. Under this generalized smoothness condition, Zhang *et al.* proved the convergence rate of gradient clipping [12] and demonstrated its advantages over gradient descent methods in neural network training. Since the introduction of this generalized smoothness condition, related research has been extensively developed in various fields, including minimax optimization [13], bilevel optimization [14], and variational inequalities [15].

On this basis, Chen *et al.* [16] further proposed a more general class of symmetric generalized smoothness conditions— α -symmetric $\left(L_0, L_1\right)$ -smoothness. This concept generalizes existing smoothness definitions and covers many mainstream machine learning problems and important function classes, such as distributionally robust optimization [17] [18], higher-order polynomials, and exponential functions. Under this framework, the researchers established corresponding descent lemmas for different values of α. On this basis, to solve non-convex optimization problems, Chen *et al.* designed several deterministic normalized gradient descent algorithms [19], which achieve an optimal complexity of $O\left(\epsilon^{-2}\right)$. Meanwhile, they found that under the stochastic setting, the classic variance reduction algorithm SPIDER [20] can also achieve the optimal sample complexity of $O\left(\epsilon^{-3}\right)$.

In the same year, Li *et al.* [21] also generalized the $\left(L_0, L_1\right)$-smoothness condition by replacing the original affine function $L_0 + L_1\|\cdot\|$ with a more general non-decreasing continuous function $\ell(\cdot)$ to characterize the dependence between the norm of the Hessian matrix and the gradient norm, thereby proposing a more generalized smoothness concept— $\ell(\cdot)$-smoothness. That is, there exists a non-decreasing continuous function $\ell(\cdot)$ such that f satisfies

$$\left\|\nabla^2 f(x)\right\| \leq \ell\left(\left\|\nabla f(x)\right\|\right).$$

This generalization significantly expands the applicability of the generalized smoothness theory, enabling researchers to handle function classes with more complex curvature variations. Based on this new framework, Li *et al.* systematically analyzed the convergence rates of gradient descent methods in convex, strongly convex, and non-convex settings. Notably, under the condition that the objective function is convex, they further proved that Nesterov's accelerated gradient method [22] achieves an optimal computational complexity of $O\left(\epsilon^{-0.5}\right)$, which exactly matches the optimal bound under the classical smoothness condition. In addition, they also conducted an in-depth study of stochastic non-convex optimization problems and, under the assumption of bounded variance, proved that stochastic gradient descent achieves an optimal complexity of $O\left(\epsilon^{-4}\right)$, which also matches its optimal bound under the classical smoothness condition.

Related Work

Overall, existing literature mainly focuses on smooth unconstrained problems,

while studies for composite optimization problems are relatively limited. In particular, for the proximal gradient method, a fundamental algorithm for solving composite optimization problems, its theoretical analysis under the generalized smooth framework remains incomplete. Although the convergence analysis of proximal gradient methods based on the Kurdyka-Łojasiewicz (KL) property has successfully extended the theoretical framework to non-convex and non-globally Lipschitz settings [8], such methods still have obvious limitations: their convergence relies on line search strategies to ensure iterative descent, and they can only yield a unified sublinear convergence rate determined by the KL exponent, failing to recover the classical exact convergence rate under convex structures. Therefore, it is necessary to re-examine the algorithmic theory for optimization problems under generalized smoothness conditions and establish a systematic analysis framework applicable to proximal gradient algorithms.

Against the above background, this paper focuses on the composite optimization problem (1) under generalized smoothness conditions, and systematically studies the convergence rate of the PGD. For problem (1), we rigorously establish the achievable convergence rate of the PGD method with a constant step size, provided that the function f is ℓ-smooth and convex. This result shows that the PGD method can still retain the same convergence rate as that under the classical Lipschitz smoothness assumption, even under the weaker generalized smoothness condition.

2. Preliminaries

This paper mainly conducts research in the finite-dimensional Euclidean space $\mathbb{R}^d$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a non-empty convex set. For any $x, y \in \mathbb{R}^d$, the inner product is denoted by $\langle x, y \rangle$, and the induced norm is defined as $\|x\| := \sqrt{\langle x, x \rangle}$. In this thesis, $B(x, R)$ denotes the Euclidean ball centered at x with radius R.

Definition 1 (ℓ-smoothness [21]). Let $f : \mathcal{X} \to \mathbb{R}$ be a real-valued differentiable function. If there exists a non-decreasing continuous function $\ell : [0, +\infty) \to (0, +\infty)$ such that the spectral norm of the Hessian matrix of f satisfies

$$\left\|\nabla^2 f(x)\right\| \le \ell\left(\left\|\nabla f(x)\right\|\right)$$

almost everywhere on $\mathcal{X}$ (with respect to the Lebesgue measure), then f is called an ℓ-smooth function.

Remark. When the function ℓ is a constant function $\ell(\mu) = L$, the definition of ℓ-smoothness degenerates to the classical smoothness definition; when ℓ is an affine function $\ell(\mu) = L_0 + L_1\mu$, the ℓ-smoothness corresponds to the (L_0, L_1)-smoothness definition [9].

Definition 2 ((r, ℓ)-smoothness). Let $f : \mathcal{X} \to \mathbb{R}$ be a real-valued differentiable function. Suppose there exist a non-decreasing continuous function $\ell : [0, +\infty) \to (0, +\infty)$ and a non-increasing continuous function $r : [0, +\infty) \to (0, +\infty)$ such that

1) $\forall x \in \mathcal{X}$, $B\left(x, r\left(\left\|\nabla f(x)\right\|\right)\right) \subseteq \mathcal{X}$;

2) $\forall x_1, x_2 \in B\left(x, r\left(\left\|\nabla f(x)\right\|\right)\right)$, $\left\|\nabla f(x_1) - \nabla f(x_2)\right\| \le \ell\left(\left\|\nabla f(x)\right\|\right)\left\|x_1 - x_2\right\|$;

Then f is called an (r,ℓ)-smooth function.

Remark. Here, the function $r:[0,+\infty)\to(0,+\infty)$ is assumed to be nonincreasing and the function $\ell:[0,+\infty)\to(0,+\infty)$ is assumed to be nondecreasing. This restriction is usually without loss of generality. When r and ℓ do not satisfy the required monotonicity, the desired monotonicity can be achieved by the following construction technique:

1) Replace the original function r with $\hat{r} := \inf_{0\le v\le u} r(v) \le r(u)$;

2) Replace the original function ℓ with the nonincreasing function $\hat{\ell} := \sup_{0\le v\le u} \ell(v) \ge \ell(u)$.

Next, we elaborate on the intrinsic relationship between the two types of smooth functions. In fact, under appropriate assumptions, ℓ-smooth functions and (r,ℓ)-smooth functions are equivalent.

Proposition 1. If f is (r,ℓ)-smooth, then f is also ℓ-smooth; if f is ℓ-smooth and f is a differentiable closed function on the open effective domain X, then f is an (r,m)-smooth function, where $m(u) := \ell(u+a)$, $r(u) := a/m(u)$, and the constant $a>0$ is arbitrary.

The proof of the above proposition can be found in appendix A.2 of [21]. The objective function considered in this paper obviously satisfies the assumptions of Proposition 1. In view of the equivalence between ℓ-smoothness and (r,ℓ)-smoothness in the above framework, the subsequent theoretical analysis will mainly focus on (r,ℓ)-generalized smooth functions.

Lemma 2 (Descent Lemma [23]). Let $f:\mathcal{X}\to\mathbb{R}$ be a M-smooth function, where $\mathcal{X}$ is a convex set. Then for all $x, y\in\mathcal{X}$, we have

$$f(y) \le f(x) + \left\langle \nabla f(x), y-x\right\rangle + \frac{M}{2}\left\|x-y\right\|^2.$$

Lemma 3. [21] Suppose the differentiable function $f:\mathcal{X}\to\mathbb{R}$ satisfies (r,ℓ)-smoothness, and for all $x\in\mathcal{X}$, $\left\|\nabla f(x)\right\| \le C$, Then

1) $B\left(x, r(C)\right) \subseteq \mathcal{X}$;

2) For any $x_1, x_2 \in B\left(x, r(C)\right)$,

$$\left\|\nabla f(x_1) - \nabla f(x_2)\right\| \le L_f \left\|x_1 - x_2\right\|, \tag{2}$$

$$f(x_1) \le f(x_2) + \left\langle \nabla f(x_2), x_1 - x_2\right\rangle + \frac{L_f}{2}\left\|x_1 - x_2\right\|^2, \tag{3}$$

where $L_f := \ell(C)$ denotes an effective smoothness constant.

Proof of the above lemma can be found in appendix A.3 of [21].

Remark. From Lemma 3, it is not difficult to see that when the gradient norm of an (r,ℓ)-smooth function is bounded, it also satisfies the property of a classical smooth function in a local neighborhood $B\left(x, r(C)\right)$, *i.e.*, the descent lemma.

Remark. From Proposition 1, the two classes of smooth functions are equivalent. Therefore, the conclusion of Lemma 3 obviously also holds for ℓ-smooth functions. When $a := C$, it suffices to take $L_f = \ell(2C)$ and $r(C) = C/L_f$ to

obtain the corresponding conclusion [21].

Lemma 4 (Nonexpansiveness of the Proximal Operator [23]). Let $g:\mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ be a closed proper convex function. Then for any $x, y \in \mathbb{R}^d$, we have

$$\left\langle x-y, \operatorname{prox}_g(x)-\operatorname{prox}_g(y)\right\rangle \geq \left\|\operatorname{prox}_g(x)-\operatorname{prox}_g(y)\right\|^2, \tag{4}$$

and

$$\left\|\operatorname{prox}_g(x)-\operatorname{prox}_g(y)\right\| \leq \|x-y\|.$$

3. The Proximal Gradient Method

3.1. The Proximal Gradient Method

This paper mainly considers the framework of generalized smoothness assumptions, and investigates the iterative convergence rate of PGD for solving unconstrained composite convex optimization problems. In the following, this chapter will discuss the convex case, present and prove the convergence rate of PGD.

Before proceeding with the study, we first introduce an important fundamental assumption in the field of optimization algorithm research, namely, the existence of an optimal solution.

Assumption 1. Assume there exists a point $x^* \in \mathcal{X}$ such that $F(x^*) = F^* = \inf_{x\in\mathcal{X}} F(x)$.

This chapter mainly considers the constant step-size version of PGD. Let the iteration step size be $\frac{1}{L}$, where $L>0$. The iteration scheme is given by

$$x^{t+1} = \operatorname{prox}_{\frac{1}{L}g}\left(x^t - \frac{1}{L}\nabla f(x^t)\right), \tag{5}$$

For simplicity, we define $T_L(x) := \operatorname{prox}_{\frac{1}{L}g}\left(x-\frac{1}{L}\nabla f(x)\right)$. Then the PGD iteration scheme is abbreviated as

$$x^{t+1} = T_L(x^t).$$

Next, we introduce an important tool, the gradient mapping, which is defined as follows.

Definition 3. When solving the objective problem (1) using PGD (5), combined with the definition of the operator T_L, the mapping $G_L:\mathcal{X}\to\mathbb{R}^d$ is defined as

$$G_L(x) = L\left(x - T_L(x)\right),\ x\in\mathcal{X},$$

and is called the gradient mapping.

Based on the above definition of the gradient mapping, the proximal gradient iteration can be equivalently transformed into a gradient descent iteration format, *i.e.*,

$$x^{t+1} = x^t - \frac{1}{L}G_L(x^t).$$

Through this structural reconstruction, it is not difficult to see that PGD can be

regarded as a natural extension of the classical gradient descent method in composite optimization problems.

Since cocoercivity serves as an important theoretical guarantee for the convergence analysis of the PGD iteration framework. However, in the composite convex optimization problem (1) considered in this paper, the function f is only (r,ℓ)-smooth. Therefore, we need to verify whether the corresponding co-coercivity still holds when the function satisfies (r,ℓ)-smoothness. Li *et al.* [21] pointed out that when a function is both convex and (r,ℓ)-smooth, its gradient also possesses cocoercivity, as stated in the following lemma.

Lemma 5 (cocoercivity [21]). Let the function $f:\mathcal{X}\to\mathbb{R}$ be an (r,ℓ)-smooth convex function. Then for $\forall x\in\mathcal{X}, y\in B\left(x,\frac{r\left(\|\nabla f(x)\|\right)}{2}\right)$, we have

$$\left\langle \nabla f(x)-\nabla f(y), x-y\right\rangle \geq \frac{1}{\ell\left(\|\nabla f(x)\|\right)}\left\|\nabla f(x)-\nabla f(y)\right\|^2. \tag{6}$$

Furthermore, we will investigate the convergence rate of PGD iterations based on the concept of co-coercivity. Since we will utilize Lemma 4 in the convergence analysis of PGD, it requires that consecutive iterates x^{k+1}, x^k of PGD satisfy $\left\|x^{k+1}-x^k\right\|=\frac{1}{L}\left\|G_L\left(x^k\right)\right\|\leq r(C)/2$, Based on this, we prove that the norm of the gradient mapping $\left\|G_L\left(x^k\right)\right\|$ is monotonically decreasing.

Lemma 6. Let $f:\mathcal{X}\to\mathbb{R}$ be an (r,ℓ)-smooth differentiable convex function satisfying $\|\nabla f(x)\|\leq C$; and let $g:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ be a proper closed convex function. If for $x\in\mathcal{X}$, $T_L(x)\in B\left(x,\frac{r(C)}{2}\right)\subseteq B\left(x,\frac{r\left(\|\nabla f(x)\|\right)}{2}\right)$, then

$$\left\|G_L\left(T_L(x)\right)\right\|\leq\left\|G_L(x)\right\|, \tag{7}$$

where $L\geq\frac{L_f}{2}=\frac{\ell(C)}{2}$.

Proof. Let $x\in\mathcal{X}, y\in B\left(x,\frac{r\left(\|\nabla f(x)\|\right)}{2}\right)$. By Lemma 4, the proximal operator is nonexpansive. When we choose $x:=x-\frac{1}{L}\nabla f(x)$, $y:=y-\frac{1}{L}\nabla f(y)$ respectively, Equation (4) becomes

$$\left\langle T_L(x)-T_L(y),\left(x-\frac{1}{L}\nabla f(x)\right)-\left(y-\frac{1}{L}\nabla f(y)\right)\right\rangle\geq\left\|T_L(x)-T_L(y)\right\|^2.$$

Since the mapping $T_L=\mathrm{Id}-\frac{1}{L}G_L$, the above inequality can be transformed into

$$\begin{aligned}&\left\langle\left(x-\frac{1}{L}G_L(x)\right)-\left(y-\frac{1}{L}G_L(y)\right),\left(x-\frac{1}{L}\nabla f(x)\right)-\left(y-\frac{1}{L}\nabla f(y)\right)\right\rangle\\&\geq\left\|\left(x-\frac{1}{L}G_L(x)\right)-\left(y-\frac{1}{L}G_L(y)\right)\right\|^2.\end{aligned}$$

Because $L>0$, the above inequality can be simplified as

$$\left\langle \left(x-\frac{1}{L}G_L(x)\right)-\left(y-\frac{1}{L}G_L(y)\right),\left(G_L(x)-\nabla f(x)\right)-\left(G_L(y)-\nabla f(y)\right)\right\rangle \geq 0,$$

Further, by expanding the inner product and rearranging terms, we obtain

$$\begin{aligned}\left\langle G_L(x)-G_L(y),x-y\right\rangle \geq{}& \left\langle \nabla f(x)-\nabla f(y),x-y\right\rangle+\frac{1}{L}\left\|G_L(x)-G_L(y)\right\|^2\\ &-\frac{1}{L}\left\langle G_L(x)-G_L(y),\nabla f(x)-\nabla f(y)\right\rangle.\end{aligned} \tag{8}$$

Since the function f is an (r,ℓ)-smooth differentiable convex function and satisfies $\|\nabla f(x)\|\leq C$, it follows from Lemma 5 that

$$\begin{aligned}\left\langle \nabla f(x)-\nabla f(y),x-y\right\rangle &\geq \frac{1}{\ell\left(\|\nabla f(x)\|\right)}\left\|\nabla f(x)-\nabla f(y)\right\|^2\\ &\geq \frac{1}{\ell(C)}\left\|\nabla f(x)-\nabla f(y)\right\|^2,\end{aligned}$$

where the second inequality holds because $\|\nabla f(x)\|\leq C$ and the function ℓ is nondecreasing and continuous. Therefore, based on the above inequality, Equation (8) can be further transformed into

$$\begin{aligned}&\ell(C)\left\langle G_L(x)-G_L(y),x-y\right\rangle\\ &\geq \frac{\ell(C)}{L}\left\|G_L(x)-G_L(y)\right\|^2+\left\|\nabla f(x)-\nabla f(y)\right\|^2\\ &\quad-\frac{\ell(C)}{L}\left\langle G_L(x)-G_L(y),\nabla f(x)-\nabla f(y)\right\rangle,\end{aligned} \tag{9}$$

For convenience, let $a:=G_L(x)-G_L(y)$, $b:=\nabla f(x)-\nabla f(y)$. Then the right-hand side of the above inequality can be further simplified as

$$\frac{\ell(C)}{L}\|a\|^2+\|b\|^2-\frac{\ell(C)}{L}a^{\mathrm{T}}b=\frac{4L\cdot\ell(C)-\ell(C)^2}{4L^2}\|a\|^2+\left\|\frac{\ell(C)}{2L}a-b\right\|^2$$

Substituting the simplified result back into (9) yields

$$\ell(C)\left\langle G_L(x)-G_L(y),x-y\right\rangle \geq \frac{4L\cdot\ell(C)-\ell(C)^2}{4L^2}\left\|G_L(x)-G_L(y)\right\|^2.$$

Since the function $\ell:[0,+\infty)\to(0,+\infty)$, we have $\ell(C)>0$ always holds. Dividing both sides of the above inequality by $\ell(C)$, we obtain

$$\left\langle G_L(x)-G_L(y),x-y\right\rangle \geq \frac{4L-\ell(C)}{4L^2}\left\|G_L(x)-G_L(y)\right\|^2. \tag{10}$$

Next, we estimate the left-hand side of (10). Using the Cauchy-Schwarz inequality, we obtain

$$\left\langle G_L(x)-G_L(y),x-y\right\rangle \leq \left\|G_L(x)-G_L(y)\right\|\|x-y\|,$$

Therefore, combining the above inequality with (10) yields

$$\left\|G_L(x)-G_L(y)\right\|\leq \frac{4L^2}{4L-\ell(C)}\|x-y\|. \tag{11}$$

According to the definition of the gradient mapping $G_L(x)=L\left(x-T_L(x)\right)$, and taking $y=T_L(x)$, we obtain

$$x-y=x-T_L(x)=\frac{1}{L}G_L(x)$$

and

$$G_L(y)=L\left(y-T_L(y)\right)=G_L\left(T_L(x)\right)=L\left(T_L(x)-T_L\left(T_L(x)\right)\right),$$

Let $x^+ \coloneqq T_L(x)$, the last equality is abbreviated as $G_L(y)=G_L\left(x^+\right)=L\left(x^+-T_L\left(x^+\right)\right)$. Substituting these two equalities into (10), we obtain

$$\begin{aligned}
&\left\langle G_L\left(x^+\right)-G_L(x),-\frac{1}{L}G_L(x)\right\rangle \geq \frac{4L-\ell(C)}{4L^2}\left\|G_L\left(x^+\right)-G_L(x)\right\|^2 \\
&\Leftrightarrow -\frac{1}{L}\left(\left\langle G_L\left(x^+\right),G_L(x)\right\rangle-\left\|G_L(x)\right\|^2\right) \\
&\quad \geq \frac{4L-\ell(C)}{4L^2}\left(\left\|G_L\left(x^+\right)\right\|^2-2\left\langle G_L\left(x^+\right),G_L(x)\right\rangle+\left\|G_L(x)\right\|^2\right) \\
&\Leftrightarrow \frac{2L-\ell(C)}{2L^2}\left\langle G_L\left(x^+\right),G_L(x)\right\rangle \geq \frac{4L-\ell(C)}{4L^2}\left\|G_L\left(x^+\right)\right\|^2-\frac{\ell(C)}{4L^2}\left\|G_L(x)\right\|^2.
\end{aligned} \tag{12}$$

1) If $L=\frac{\ell(C)}{2}$, Equation (12) can be simplified as

$$\frac{1}{\ell(C)}\left\|G_L(x)\right\|^2 \geq \frac{1}{\ell(C)}\left\|G_L\left(x^+\right)\right\|^2,$$

Multiplying both sides by $\ell(C)>0$, we obtain

$$\left\|G_L(x)\right\|^2 \geq \left\|G_L\left(x^+\right)\right\|^2.$$

Further, by letting $x^+ \coloneqq T_L(x)$, we arrive at the final conclusion $\left\|G_L(x)\right\|^2 \geq \left\|G_L\left(T_L(x)\right)\right\|^2$.

2) Similarly, when $L\in\left(\frac{\ell(C)}{2},+\infty\right)$, combining with Equation (12) and using the Cauchy-Schwarz inequality, we obtain

$$\frac{2L-\ell(C)}{2L^2}\left\|G_L\left(x^+\right)\right\|\left\|G_L(x)\right\| \geq \frac{4L-\ell(C)}{4L^2}\left\|G_L\left(x^+\right)\right\|^2-\frac{\ell(C)}{4L^2}\left\|G_L(x)\right\|^2, \tag{13}$$

By rearranging terms and factoring, the above inequality becomes

$$\left(\left\|G_L(x)\right\|-\left\|G_L\left(x^+\right)\right\|\right)\left(\frac{\ell(C)}{4L^2}\left\|G_L(x)\right\|+\frac{4L-\ell(C)}{4L^2}\left\|G_L\left(x^+\right)\right\|\right)\geq 0. \tag{14}$$

(i) If $\left\|G_L\left(x^+\right)\right\|=0$, by the non-negativity of the norm we have $\left\|G_L(x)\right\|\geq 0$. Furthermore, since $\left\|G_L\left(x^+\right)\right\|=0$, it follows that $\left\|G_L(x)\right\|\geq\left\|G_L\left(x^+\right)\right\|$, the conclusion holds.

(ii) If $\left\|G_L\left(x^+\right)\right\|\neq 0$, by the non-negativity of the norm we have $\left\|G_L\left(x^+\right)\right\|>0$.

In this case, $\|G_L(x)\|>0$ must also hold. If $\|G_L(x)\|>0$ were not true, then by the non-negativity of the norm we would have $\|G_L(x)\|=0$. From (11), with $x^+ := T_L(x)$ and $G_L(x)=L(x-T_L(x))$, we obtain

$$\begin{aligned}\|G_L(x^+)-G_L(x)\| &\le \frac{4L^2}{4L-\ell(C)}\|x^+-x\| = \frac{4L^2}{4L-\ell(C)}\|T_L(x)-x\| \\ &= \frac{4L^2}{4L-\ell(C)}\cdot\frac{1}{L}\|G_L(x)\|,\end{aligned}$$

Since $\|G_L(x)\|=0$, it follows that $\|G_L(x^+)\|\le 0$, which contradicts the premise $\|G_L(x^+)\|>0$. Therefore, given $L>\frac{\ell(C)}{2}$ and $\ell(C)>0$, we have $4L-\ell(C)>0$, and consequently

$$\frac{\ell(C)}{4L^2}\|G_L(x)\|+\frac{4L-\ell(C)}{4L^2}\|G_L(x^+)\|>0.$$

At this point, combining the above inequality with (14), it is easy to see that $\|G_L(x)\|\ge\|G_L(x^+)\|$, so the conclusion holds.

□

Next, we present the key descent inequality in the convergence analysis.

Lemma 7 (Basic Proximal Gradient Inequality). Let $f:\mathcal{X}\to\mathbb{R}$ be a differentiable convex function, where $\mathcal{X}$ is an open set, and let $g:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ be a proper closed convex function. For any $x,y\in\mathcal{X}$ and $L>0$, if

$$f(T_L(x))\le f(y)+\langle\nabla f(y),T_L(y)-y\rangle+\frac{L}{2}\|T_L(y)-y\|^2, \tag{15}$$

then

$$F(x)-F(T_L(y))\ge\frac{L}{2}\|x-T_L(y)\|^2-\frac{L}{2}\|x-y\|^2+h(x,y), \tag{16}$$

where $h(x,y)=f(x)-f(y)-\langle\nabla f(y),x-y\rangle$.

Proof. Define the function

$$\phi(u) := f(y)+\langle\nabla f(y),u-y\rangle+g(u)+\frac{L}{2}\|u-y\|^2. \tag{17}$$

Clearly, ϕ is a strongly convex function with strong convexity modulus L. From the definition of $T_L(x)$, we have

$$\begin{aligned}T_L(y) &= \operatorname{prox}_{\frac{1}{L}g}\left(y-\frac{1}{L}\nabla f(y)\right) \\ &= \arg\min_{u\in\mathbb{R}^d}\left\{g(u)+\frac{L}{2}\left\|u-\left(y-\frac{1}{L}\nabla f(y)\right)\right\|^2\right\} \\ &= \arg\min_{u\in\mathbb{R}^d}\phi(u).\end{aligned} \tag{18}$$

By strong convexity, we obtain

$$\phi(x)-\phi(T_L(y))\ge\frac{L}{2}\|x-T_L(y)\|^2. \tag{19}$$

Using the convexity of f, we have

$$f(y) \geq f(T_L(y)) + \langle \nabla f(y), y - T_L(y) \rangle.$$

Letting $u := T_L(y)$ and substituting it into (17), together with the above inequality, we get

$$\begin{aligned} \phi(T_L(y)) &= f(y) + \langle \nabla f(y), T_L(y) - y \rangle + \frac{L}{2}\|T_L(y) - y\|^2 + g(T_L(y)) \\ &\geq f(T_L(y)) + g(T_L(y)) = F(T_L(y)), \end{aligned}$$

furthermore, from (19), it is not difficult to see that

$$\phi(x) - F(T_L(x)) \geq \frac{L}{2}\|x - T_L(y)\|^2.$$

Next, substituting $u := x$ into the expression for $\phi(u)$ and combining with the above inequality, we obtain

$$f(y) + \langle \nabla f(x), x - y \rangle + g(x) + \frac{L}{2}\|x - y\|^2 - F(T_L(y)) \geq \frac{L}{2}\|x - T_L(y)\|^2,$$

by rearranging terms, we get

$$F(x) - F(T_L(y)) \geq \frac{L}{2}\|x - T_L(y)\|^2 - \frac{L}{2}\|x - y\|^2 + f(x) - f(y) - \langle \nabla f(x), x - y \rangle.$$

This completes the proof. □

Since y is arbitrary, we may replace y with x, and thus the following corollary can be deduced from Lemma 7.

Corollary 1. Let $f : \mathcal{X} \to \mathbb{R}$ be a continuously differentiable convex function, where $\mathcal{X}$ is an open set, and let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper closed convex function. For any $x, y \in \mathcal{X}$ and $L > 0$, if

$$f(T_L(x)) \leq f(x) + \langle \nabla f(x), T_L(x) - x \rangle + \frac{L}{2}\|T_L(x) - x\|^2,$$

then

$$F(x) - F(T_L(x)) \geq \frac{1}{2L}\|G_L(x)\|^2.$$

Since the consecutive iterates x^{t+1} and x^t of the proximal gradient method satisfy $x^{t+1} = T_L(x^t)$, Corollary 1 directly establishes the sufficient descent property of the PGD algorithm. In what follows, we combine the above lemmas to study the convergence rate of PGD for solving convex problems.

3.2. Convergence Rate Analysis of the Proximal Gradient Method in the Convex Case

In this subsection, we consider the case where f in the objective function is convex, and analyze the convergence of PGD iterations under the generalized smoothness assumption. Specifically, the convergence rate results obtained through theoretical analysis are as follows.

Theorem 8. Consider problem (1). Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper closed convex function, and let $f : \mathcal{X} \to \mathbb{R}$ be a differentiable proper closed convex

function, where $\mathcal{X}$ is an open convex set and $\operatorname{dom} g \subseteq \mathcal{X}$. Assume that f is an (r,ℓ)-smooth function, where $\ell:[0,+\infty)\to(0,+\infty)$ is a nondecreasing continuous function and $r:[0,+\infty)\to(0,+\infty)$ is a nonincreasing continuous function. Moreover, suppose that f is C-Lipschitz continuous. If the parameter L satisfies

$$L \geq \max\left\{\ell(C), 2\left\|G_L(x_0)\right\|/r(C)\right\},$$

then

$$F\left(x^k\right)-F^* \leq \frac{L\left\|x^0-x^*\right\|^2}{2k}.$$

Proof. Since f is C-Lipschitz continuous, for any $x, y \in \mathcal{X}$, we have

$$\left\|\nabla f(x)\right\| = \lim_{y\to x}\frac{\left\|f(y)-f(x)\right\|}{|y-x|} \leq \frac{C|y-x|}{|y-x|} = C.$$

Given that f is convex and (r,ℓ)-smooth, and g is a proper closed convex function. First, from

$$x^{t+1} = T_L\left(x^t\right)$$

and $G_L(x) = L\left(x - T_L(x)\right)$, together with Lemma 6, we obtain

$$\left\|x^{t+1}-x^t\right\| = \left\|T_L\left(x^t\right)-x^t\right\| = \frac{1}{L}\left\|G_L\left(x^t\right)\right\| \leq \frac{1}{L}\left\|G_L\left(x^{t-1}\right)\right\| \leq \cdots \leq \frac{1}{L}\left\|G_L\left(x^0\right)\right\| \leq \frac{r(C)}{2},$$

which implies

$$x^{t+1} \in B\left(x^t, \frac{r(C)}{2}\right).$$

Secondly, by Lemma 3, the two consecutive iterates x^{t+1}, x^t of PGD always satisfy

$$f\left(x^{t+1}\right) \leq f\left(x^t\right) + \left\langle \nabla f\left(x^t\right), x^{t+1}-x^t\right\rangle + \frac{\ell(C)}{2}\left\|x^{t+1}-x^t\right\|^2. \tag{20}$$

Finally, in Lemma 7, due to the arbitrariness of L, we may set $L = \ell(C)$. Together with $x^{t+1} = T_L\left(x^t\right)$, we have for any $t \in \mathbb{N}^*$,

$$\frac{2}{\ell(C)}\left(F\left(x^*\right)-F\left(x^{t+1}\right)\right) \geq \left\|x^*-x^{t+1}\right\|^2 - \left\|x^*-x^t\right\|^2 + \frac{2}{\ell(C)}h\left(x^*, x^t\right), \tag{21}$$

where, by Assumption 1, x^* is the optimal solution of the objective function, and the second inequality follows from the convexity of f. Now, letting t run over $0,1,2,\cdots,k-1$ and summing the inequalities, we obtain

$$\frac{2}{\ell(C)}\sum_{t=0}^{k-1}\left(F\left(x^*\right)-F\left(x^{t+1}\right)\right) \geq \left\|x^*-x^t\right\|^2 - \left\|x^*-x^0\right\|^2.$$

Multiplying both sides of the above inequality by $\frac{\ell(C)}{2}$, we get

$$\sum_{t=0}^{k-1}\left(F\left(x^*\right)-F\left(x^t\right)\right) \leq \frac{\ell(C)}{2}\left\|x^*-x^0\right\|^2 - \frac{\ell(C)}{2}\left\|x^*-x^k\right\|^2 \leq \frac{\ell(C)}{2}\left\|x^*-x^0\right\|^2.$$

From Corollary 1, we also have $F\left(x^{t+1}\right) \le F\left(x^*\right)$. Combining this with the above inequality yields

$$\frac{2}{\ell(C)}\sum_{t=0}^{k-1}\left(F\left(x^*\right)-F\left(x^{t+1}\right)\right)^2 \ge \left\|x^*-x^k\right\|^2-\left\|x^*-x^0\right\|^2.$$

Multiplying both sides of the above inequality by $\frac{\ell(C)}{2}$, we obtain

$$\sum_{t=0}^{k-1}\left(F\left(x^*\right)-F\left(x^t\right)\right) \le \frac{\ell(C)}{2}\left\|x^*-x^0\right\|^2-\frac{\ell(C)}{2}\left\|x^*-x^k\right\|^2 \le \frac{\ell(C)}{2}\left\|x^*-x^0\right\|^2.$$

From Corollary 1, we also have $F\left(x^{t+1}\right) \le F\left(x^t\right)$. Combining this with the above inequality yields

$$k\left(F\left(x^k\right)-F\left(x^*\right)\right) \le \sum_{t=1}^{k-1}\left(F\left(x^{t+1}\right)-F\left(x^*\right)\right) \le \frac{\ell(C)}{2}\left\|x^*-x^0\right\|^2,$$

which can be rearranged as

$$F\left(x^k\right)-F\left(x^*\right) \le \frac{\ell(C)\left\|x^*-x^0\right\|^2}{2k} \le \frac{L\left\|x^*-x^0\right\|^2}{2k}.$$

This completes the proof. □

4. Conclusions

To address the strict limitation of the classic gradient Lipschitz smoothness assumption on the applicability of the PGD, this paper introduces a novel class of generalized smoothness conditions— $\ell(\cdot)$ —smoothness, establishing a more general framework for analyzing composite convex optimization problems. Under this generalized smoothness framework, we systematically derive the convergence theory of the constant-stepsize PGD algorithm and rigorously prove that: even under the generalized smoothness assumption that relaxes the classic Lipschitz gradient continuity, the constant-stepsize PGD can still achieve the exact same sublinear convergence rate $O(1/k)$ as in the classic smoothness scenario, provided that the function f satisfies $\ell(\cdot)$ —smoothness and Lipschitz continuity.

However, the analysis framework in this paper relies on the core assumption that the objective function f satisfies gradient Lipschitz continuity. While this assumption guarantees the gradient Lipschitz continuity of f in local regions and provides theoretical support for the convergence analysis of the constant-step-size PGD algorithm, it also significantly limits the scope of application of the method: for composite convex optimization problems that do not satisfy gradient Lipschitz continuity (such as scenarios with unbounded gradients or rapidly changing gradients), the sublinear convergence rate theory of the constant-stepsize PGD algorithm established in this paper will no longer hold.

Acknowledgements

Sincere thanks to the members of JAMP for their professional performance, and special thanks to managing editor *Hellen XU* for a rare attitude of high quality.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Shechtman, Y., Beck, A. and Eldar, Y.C. (2014) GESPAR: Efficient Phase Retrieval of Sparse Signals. *IEEE Transactions on Signal Processing*, **62**, 928-938. https://doi.org/10.1109/tsp.2013.2297687

[2] Cai, J.F., Long, Y., Wen, R.X. and Ying, J. (2023) A Fast and Provable Algorithm for Sparse Phase Retrieval. https://doi.org/10.48550/arXiv.2309.02046

[3] Bai, Y., Wang, Y.X. and Liberty, E. (2019) ProxQuant: Quantized Neural Networks Via Proximal Operators. 7*th International Conference on Learning Representations* (*ICLR*), New Orleans, 6-9 May 2019, 1-20.

[4] Vũ, B.C. (2013) A Splitting Algorithm for Dual Monotone Inclusions Involving Co-coercive Operators. *Advances in Computational Mathematics*, **38**, 667-681. https://doi.org/10.1007/s10444-011-9254-8

[5] Rockafellar, R.T. (2001) Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, **14**, 877-898. https://doi.org/10.1137/0314056

[6] Faw, M., Tziotis, I., Caramanis, C. and Mokhtari, A. (2022) The Power of Adaptivity in SGD: Self-Tuning Step Sizes with Unbounded Gradients and Affine Variance. *Proceedings of the* 35*th Conference on Learning Theory* (*COLT*), London, 2-5 July 2022, 313-355.

[7] De Marchi, A., Jia, X., Kanzow, C. and Mehlitz, P. (2023) Constrained Composite Optimization and Augmented Lagrangian Methods. *Mathematical Programming*, **201**, 863-896. https://doi.org/10.1007/s10107-022-01922-4

[8] Jia, X., Kanzow, C. and Mehlitz, P. (2023) Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka-Łojasiewicz Property without Global Lipschitz Assumptions. *SIAM Journal on Optimization*, **33**, 3038-3056. https://doi.org/10.1137/23m1548293

[9] Zhang, J., He, T., Sra, S. and Jadbabaie, A. (2020) Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. https://doi.org/10.48550/arXiv.1905.11881

[10] Merity, S., Keskar, N.S. and Socher, R. (2018) Regularizing and Optimizing LSTM Language Models. https://doi.org/10.48550/arXiv.1708.02182

[11] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[12] Mikolov, T. (2012) Statistical Language Models Based on Neural Networks. Brno University of Technology, Faculty of Information Technology. https://www.fit.vut.cz/study/phd-thesis/283/

[13] Xian, W., Chen, Z. and Huang, H. (2024) Delving into the Convergence of Generalized Smooth Minimax Optimization. *Communications in Mathematical Physics*, **235**, 54191-54211.

[14] Hao, J., Gong, X. and Liu M. (2024) Bilevel Optimization under Unbounded Smoothness: A New Algorithm and Convergence Analysis. https://arxiv.org/abs/2401.09587

[15] Vankov, D., Nedich, A. and Sankar, L. (2024) Generalized Smooth Variational Inequalities: Methods with Adaptive Stepsizes. *Proceedings of the* 41*st International*

Conference on Machine Learning (*ICML*), Vienna, 21-27 July 2024, 49137-49170. https://proceedings.mlr.press/v235/vankov24a.html

[16] Chen, Z., Zhou, Y., Liang, Y. and Lu, Z. (2023) Generalized-Smooth Nonconvex Optimization Is as Efficient as Smooth Nonconvex Optimization. *Proceedings of the 40th International Conference on Machine Learning* (*ICML*), Honolulu, 23-29 July 2023, 5396-5427. https://proceedings.mlr.press/v202/chen23v.html

[17] Levy, D., Carmon, Y., Duchi, J.C. and Sidford, A. (2020) Large-Scale Methods for Distributionally Robust Optimization. https://arxiv.org/abs/2010.05893

[18] Jin, J., Zhang, B., Wang, H. and Wang, L. (2021) Non-Convex Distributionally Robust Optimization: Non-Asymptotic Analysis. https://arxiv.org/abs/2110.12459

[19] Cortés, J. (2006) Finite-time Convergent Gradient Flows with Applications to Network Consensus. *Automatica*, **42**, 1993-2000. https://doi.org/10.1016/j.automatica.2006.06.015

[20] Fang, C., Li, C., Lin, Z. and Zhang, T. (2018) SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator. https://arxiv.org/pdf/1807.01695

[21] Li, H., Tian, Y., Rakhlin, A. and Jadbabaie, A. (2023) Convex and Non-Convex Optimization under Generalized Smoothness. https://arxiv.org/pdf/2306.01264.pdf

[22] Nesterov, Y.E. (1983) A Method for Solving the Convex Programming Problem with Convergence Rate O($1/k^2$). *Proceedings of the USSR Academy of Sciences*, **269**, 543-547.

[23] Beck, A. (2017) First-Order Methods in Optimization. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974997