PROJECTIVE PROXIMAL GRADIENT DESCENT FOR NONCONVEX NONSMOOTH OPTIMIZATION: FAST CONVERGENCE WITHOUT KURDYKA-LOJASIEWICZ (KL) PROPERTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonconvex and nonsmooth optimization problems are important and challenging for statistics and machine learning. In this paper, we propose Projected Proximal Gradient Descent (PPGD) which solves a class of nonconvex and nonsmooth optimization problems, where the nonconvexity nd nonsmoothness come from a nonsmooth regularization term which is nonconvex but piecewise convex. In contrast with existing convergence analysis of accelerated PGD methods for nonconvex and nonsmooth problems based on the Kurdyka-Lojasiewicz (KL) property, we provide a new theoretical analysis showing that PPGD achieves optimal convergence rate on a class of nonconvex and nonsmooth problems under mild assumptions, which is the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continuous gradient. Experimental results demonstrate the effectiveness of the PPGD.

1 INTRODUCTION

Nonconvex and nonsmooth optimization problems are challenging ones which have received a lot of attention in statistics and machine learning (Bolte et al., 2014; Ochs et al., 2015). In this paper, we consider fast optimization algorithms for a class of nonconvex and nonsmooth problems presented as

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}), \tag{1}$$

where $h(\mathbf{x}) = \sum_{j=1}^{d} h_j(\mathbf{x}_j)$ is a separable regularizer, each h_j is piecewise convex, and $\inf_{x\in\mathbb{R}^d} F(x) > -\infty$. A piecewise convex function is defined in Definition 1.1. Without loss of generality we let $h_j = f$ for all $j \in [d]$, and f is a piecewise convex function. f can be either nonconvex or convex, and all the results in this paper can be straightforwardly extended to the case when $\{h_j\}$ are different.

Definition 1.1. A univariate function $f\colon \mathbb{R} \to \mathbb{R}$ is piecewise convex if f is lower semicontinuous and there exist intervals $\{\mathcal{R}_m\}_{m=1}^M$ such that $\mathbb{R} = \bigcup_{m=1}^M \mathcal{R}_m$, and f restricted on $\mathcal{R}_m$ is convex for each $m \in [M]$. The left and right endpoints of $\mathcal{R}_m$ are denoted by q_{m-1} and q_m for all $m \in [M]$, where $\{q_m\}_{m=0}^M$ are the endpoints such that $q_0 = -\infty \leq q_1 < q_2 < \ldots < q_M = +\infty$. Furthermore, f is either left continuous or right continuous at each endpoint q_m for $m \in [M-1]$. $\{\mathcal{R}_m\}_{m=1}^M$ are also referred to as convex pieces throughout this paper.

It is important to note that for $m \in [M-1]$, when f is continuous at some endpoint q_m for $m \in [M-1]$, then $q_m \in \mathcal{R}_m \cap \mathcal{R}_{m+1}$. If f is only left continuous at q_m, $q_m \in \mathcal{R}_m$ and $q_m \notin \mathcal{R}_{m+1}$. If f is only right continuous at q_m, $q_m \notin \mathcal{R}_m$ and $q_m \in \mathcal{R}_{m+1}$.

When $M = 1$, f becomes a convex function on $\mathbb{R}$, and problem (1) is a convex problem. We consider $M > 1$ throughout this paper, and our proposed algorithm trivially extends to the case when $M = 1$. We allow a special case that an interval $\mathcal{R}_m = \{q_m\}$ for $m \in [M-1]$ is a single-point set, in this case $q_{m-1} = q_m$. When there are no single-point intervals in $\{\mathcal{R}_m\}_{m=1}^M$, the

minimum interval length is defined by $R_0 = \min_{m\in[M-1]} |\mathcal{R}_m|$. Otherwise,the minimum interval length is $R_0 = \min_{m\in[M-1]:\, |\mathcal{R}_m|\neq 0} |\mathcal{R}_m|$, where $|\cdot|$ denotes the length of an interval.

Below are three examples of piecewise convex functions with the corresponding regularizers which have been widely used in constrained optimization and sparse learning problems.

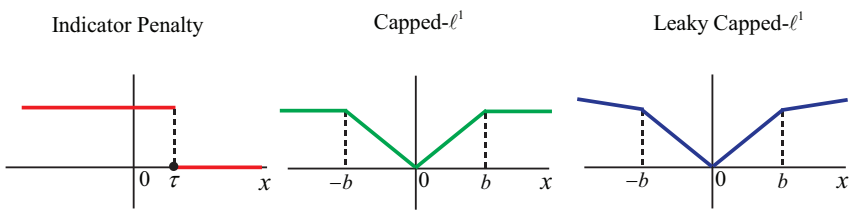


Figure 1: Illustration of three piecewise convex functions.

Example 1.1. (1) The indicator penalty function $f(x) = \lambda \mathbb{1}_{\{x<\tau\}}$ is piecewise convex with $\mathcal{R}_1 = (-\infty, \tau), \mathcal{R}_2 = [\tau, \infty)$. (2) The capped-$\ell^1$ penalty function $f(x) = f(x; \lambda, b) = \lambda \min\{|x|, b\}$ is piecewise convex with $\mathcal{R}_1 = (-\infty, -b], \mathcal{R}_2 = [-b, b], \mathcal{R}_3 = [b, \infty)$. (3) The leaky capped-$\ell^1$ penalty function (Wangni & Lin, 2017) $f = \lambda \min\{|x|, b\} + \beta \mathbb{1}_{\{|x|\geq b\}} |x - b|$ with $\mathcal{R}_1 = (-\infty, -b], \mathcal{R}_2 = [-b, b], \mathcal{R}_3 = [b, \infty)$. The three functions are illustrated in Figure 1. While not illustrated, $f(x) = \mathbb{1}_{\{x\neq 0\}}$ for the ℓ^0-norm with $h(\mathbf{x}) = \|\mathbf{x}\|_0$ is also piecewise convex.

A piecewise convex function f on each of its convex piece $\mathcal{R}_m$ can be viewed as an restriction of a globally convex function f_i to that convex piece. That is, there exists a convex function $f_i : \mathbb{R} \to \mathbb{R}$ such that $f_i(x) = f(x)$ for all $i \in [m]$ and $x \in \mathcal{R}_m$.

The main assumption of this paper is that g and h satisfy the following conditions.

Assumption 1 (Main Assumption)**.** (a) g is convex with L-smooth gradient, that is, $\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|_2 \leq L_g \|\mathbf{x} - \mathbf{y}\|_2$. g has bounded gradient, and $\|\nabla g(\mathbf{x})\|_\infty \leq G$ for a positive constant G for all $\mathbf{x} \in \mathbb{R}^d$.

(b) $f : \mathbb{R} \to \mathbb{R}$ is a piecewise convex function. Furthermore, f satisfies the following two conditions. (1) There exists a small positive constant $s_0 < R_0$ such that f is differential on $(q_m - s_0, q_m)$ and $(q_m, q_m + s_0)$ for all $m \in [M - 1]$.

(c) The proximal mapping prox_{sf_i} has closed form solution for all $m \in [-M, M]$, where $\mathsf{prox}_{sf_i}(x) := \mathrm{argmin}_{v\in\mathbb{R}} \frac{1}{2s}(v - x)^2 + f_i(v)$.

(d) f has "negative curvature" at each endpoint q_m where f is continuous at for all $m \in [M - 1]$, that is, $\lim_{x\to q_m^-} f'(x) > \lim_{x\to q_m^+} f'(x)$. We define

$$C := \min_{m\in[M-1]:\ f \text{ continuous at } q_m} \left\{ \lim_{x\to q_m^-} f'(x) - \lim_{x\to q_m^+} f'(x) \right\} > 0. \tag{2}$$

In addition, f has bounded Fréchet subdifferential, that is, $\sup_{x\in\bigcup_{m=1}^M \mathcal{R}_m^o} \sup_{v\in\tilde{\partial} f(x)} \|v\|_2 \leq F$ for some absolute constant $F > 0$, where $\mathcal{R}^o$ denotes the interior of an interval.

Fréchet subdifferential is formally defined in Section 3. It is noted that on each $\mathcal{R}_m^o$, the convex differential of f coincides with its Fréchet subdifferential. We define the minimum jump value of f at noncontinuous endpoints by

$$J := \min \left\{ \min_{\substack{m\in[M-1]: \\ f \text{ is only right continuous at } q_m}} \left\{ \left| \lim_{y\to q_m^-} f(y) - f(q_m) \right| \right\}, \min_{\substack{m\in[M-1]: \\ f \text{ is only left continuous at } q_m}} \left\{ \left| \lim_{y\to q_m^+} f(y) - f(q_m) \right| \right\} \right\}. \tag{3}$$

Assumption 1 is mild and explained as follows. When g is a G-Lipschitz continuous function, its gradient is bounded by G, and Assumption 1(a) holds. Smooth gradient is a commonly required in the standard analysis of proximal methods for non-convex problems, such as (Bolte et al., 2014;

Ghadimi & Lan, 2016a). In addition, Assumption 1(b)-(d) are mild, and they hold for all the three piecewise convex functions in Example 1.1 as well as the ℓ^0-norm.

It is noted that (1) covers a broad range of optimization problems in machine learning and statistics. The nonnegative programming problem $\min_{\mathbf{x}\in\mathbb{R}^d:\mathbf{x}_i\geq 0, i\in[d]} g(\mathbf{x})$ can be reduced to the regularized problem $g(\mathbf{x}) + h(\mathbf{x})$ with f being the indicator penalty $f(x) = \lambda \mathbb{I}_{\{x<\tau\}}$ for a properly large λ. When g is the squared loss, that is, $g(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2$ where $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{D} \in \mathbb{R}^{n\times d}$ is the design matrix, (1) is the well-known regularized sparse linear regression problems with convex or nonconvex regularizers such as the capped-ℓ^1 or ℓ^0-norm penalty.

We propose a novel method termed Projective Proximal Gradient Descent (PPGD) to solve problem (1), and PPGD enjoys fast convergence rate by a novel projection operator and a negative curvature exploitation procedure. Our main results are summarized in the subsection below.

1.1 Main Results and Contributions

We extend the existing Accelerated Proximal Gradient descent method (APG) (Beck & Teboulle, 2009b; Nesterov, 2013) on problem (1), showing improved convergence rates. More concretely,

1. We propose Projected Proximal Gradient Descent (PPGD). Using the novel and carefully designed projection operator and the Negative-Curvature-Exploitation algorithm (Algorithm 3), PPGD achieves a fast convergence rate for the nonconvex and nonsmooth optimization problem (1) which matches the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continuous gradient. In particular, it is proved that under two mild assumptions, Assumption 1 and Assumption 2 to be introduced in Section 3, there exists a finite $k_0 \geq 1$ such that for all $k > k_0$,
$$\|F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\|_2 \leq \mathcal{O}(\frac{1}{k^2}). \tag{4}$$
Details are deferred to Theorem 3.5 in Section 3. It should be emphasized that this is the same convergence rate as that of regular APG on convex problems (Beck & Teboulle, 2009a;b).
2. Our analysis provides insights into accelerated PGD methods for challenging nonconvex and nonsmooth problems. In contrast to most existing accelerated PGD methods (Li & Lin, 2015; Li et al., 2017) which employ the Kurdyka-Lojasiewicz (KL) property to analyze the convergence rates, PPGD provides an new perspective of convergence analysis without the KL property while enjoying the Nesterov's optimal convergence rate. Our analysis reveals that the objective function makes progress, that is, its value is decreased by a positive amount, when the variable sequence generated by PPGD transits from one convex piece to another. Such observation opens the door for future research in analyzing convergence rate of accelerated APG methods without the KL property.

1.2 Related Works

Assuming that both g and h are The accelerated PGD algorithms in (Li & Lin, 2015; Li et al., 2017) have linear convergence rate with $\theta \in [1/2, 1)$, and sublinear rate $\mathcal{O}(k^{-\frac{1}{1-2\theta}})$ with $\theta \in (0, 1/2)$, where θ is the Kurdyka-Lojasiewicz (KL) exponent and both rates are for objective values. The KL property of proper and lower semicontinuous functions is defined below.

Definition 1.2 (Kurdyka-Lojasiewicz (KL) Property)**.** Let $u\colon \mathbb{R}^d \to (-\infty, \infty]$ be proper and lower semicontinuous. u is said to satisfy the Kurdyka-Lojasiewicz (KL) property at $\mathbf{u} \in \operatorname{dom}\partial u := \{\mathbf{u}' \in \mathbb{R}^d\colon \partial u(\mathbf{u}') \neq \emptyset\}$, if there exist $\eta \in (0, \infty]$, a neighborhood U of $\mathbf{u}$ and a function $\phi \in \Phi_\eta$, such that for all $\bar{\mathbf{u}} \in U \cup [u(\mathbf{u}) < u(\bar{\mathbf{u}}) < u(\mathbf{u}) + \eta]$, the following inequality holds:
$$\phi'\big(u(\bar{\mathbf{u}}) - u(\mathbf{u})\big)\operatorname{dist}(\mathbf{0}, \partial u(\bar{\mathbf{u}})) \geq 1.$$
Here $\eta \in (0, \infty]$, Φ_η is defined to be the class of all concave and continuous functions $\phi\colon [0, \eta) \to \mathbb{R}^+$ which satisfy the following conditions:

1. $\phi(0) = 0$;
2. ϕ is continuously differentiable on $(0, \eta)$ and continuous at 0;

3. $\phi'(r) > 0$ for all $r \in (0, \eta)$.

If u satisfies the KL property at each point of $\operatorname{dom} \partial u$, then u is called a KL function.

To the best of our knowledge, most existing convergence analysis of accelerated PGD methods for nonconvex and nonsmooth problems, where the nonconvexity and nonsmoothness are due to the regularizer h, are based on the KL property. While (Ghadimi & Lan, 2016b) provides analysis of an accelerated PGD method for nonconvex and nonsmooth problems, the nonconvexity comes from the smooth part of the objective function (g in our objective function), so it cannot handle the problems discussed in this paper. Furtheremore, the fast PGD method by (Yang & Yu, 2020) is restricted to ℓ^0-regularized problems. As a result, it remains an interesting and important open problem regarding the convergence rate of accelerated PGD methods for problem (1).

Another line of related works focuses on accelerated gradient methods for smooth objective functions. For example, (Jin et al., 2018) proposes perturbed accelerated gradient method which achieves faster convergence by decreasing a quantity named Hamiltonian, which is the sum of the current objective value and the scaled squared distance between consecutive iterates, by exploiting the negative curvature. The very recent work (Li & Lin, 2022) further removes the polylogarithmic factor in the time complexity of (Jin et al., 2018) by a restarting strategy.

Because our objective is nonsmooth, the Hamiltonian cannot be decreased by exploiting the negative curvature of the objective in the same way as (Jin et al., 2018). However, we still manage to design a "Negative-Curvature-Exploitation" algorithm which decreases the objective value by an absolute positive amount when the iterates generated by PPGD cross endpoints. Our results are among the very few results in the optimization literature about fast optimization methods for nonconvex and nonsmooth problems which are not limited to ℓ^0-regularized problems.

1.3 Notations

Throughout this paper, we use bold letters for matrices and vectors, regular lower letters for scalars. The bold letter with subscript indicates the corresponding element of a matrix or vector. $\|\cdot\|_p$ denotes the ℓ_p-norm of a vector, or the p-norm of a matrix. $|\mathbf{A}|$ denotes the cardinality of a set $\mathbf{A}$. $[n]$ is the set of natural numbers between 1 and n inclusively. When $\mathcal{R} \subseteq \mathbb{R}$ is an interval, $\overline{\mathcal{R}}$ is the closure of $\mathcal{R}$, $\mathcal{R}^+$ denotes the set of all the points in $\mathbb{R}$ to the right of $\mathcal{R}$ while not in $\mathcal{R}$, and $\mathcal{R}^-$ is defined similarly denoting the set of all the points to the left of $\mathcal{R}$. The domain of any function u is denoted by $\mathbf{dom} u$. $\lim_{x\to a^+}$ and $\lim_{x\to a^-}$ denote left limit and right limit at a point $a \in \mathbb{R}$. $\mathbb{N}$ is the set of all natural numbers.

2 ALGORITHMS

Before presenting the proposed algorithm, we first present the definition of proximal mapping, and then describe how to build surrogate functions for each convex piece. These surrogate functions are used in the proposed PPGD algorithm.

Definition 2.1 (Proximal Mapping)**.** The proximal mapping associated with function u is defined as $\mathsf{prox}_u(\mathbf{x}) := \operatorname{argmin}_{\mathbf{v}\in\mathbb{R}^d} u(\mathbf{v}) + \frac{1}{2}\|\mathbf{v}-\mathbf{x}\|_2^2$ for $\mathbf{x} \in \mathbb{R}^d$, and $\mathsf{prox}_u(\cdot)$ is called the proximal mapping associated with u.

2.1 Construction of the Surrogate Functions $\{f_m\}_{m=1}^M$

Given that f is convex on each convex piece $\mathcal{R}_m$, we describe how to construct a surrogate function $f_m \colon \mathbb{R} \to \mathbb{R}$ such that $f_m(x) = f(x)$ for all $x \in \mathcal{R}_m$. If the right endpoint $q = q_m$ is not $+\infty$ and f is continuous at q, then f_m extends $f|_{\mathcal{R}_m}$ such that f_m on $(q, +\infty)$ is linear. Similar extension applies to the case when $q = q_{m-1}$ is the right endpoint of $\mathcal{R}_m$. Formally, we let

$$f_m(x) = \begin{cases} f(q_m) + v^-(x - q_m) & f \text{ is continuous at } q_m, \\ \lim_{y\to q_m^-} f(y) + v^-(x - q_m) & f(q_m) < \lim_{y\to q_m^-} f(y) \\ \lim_{y\to q_m^+} f(y) & f(q_m) < \lim_{y\to q_m^+} f(y), \end{cases} \tag{5}$$

for all $x \in \mathcal{R}_m^+$ if $q_m \neq +\infty$, where $v^- = \lim_{x\to q_m^-} f'(x)$. Note that when $\lim_{y\to q_m^+} f(y) > f(q_m)$, f is not continuous at q_m. According to Assumption 1, f is either left continuous or right continuous

at q_m. By the lower semicontinuity of f, it follows that f must be left continuous at $q_m \in \mathcal{R}_m$. The surrogate function f_m is defined similarly on $\mathcal{R}_m^-$ by

$$f_m(x) = \begin{cases} f(q_{m-1}) + v^+(x - q_{m-1}) & f \text{ is continuous at } q_{m-1}, \\ \lim_{y \to q_{m-1}^+} f(y) + v^+(x - q_{m-1}) & f(q_{m-1}) < \lim_{y \to q_{m-1}^+} f(y) \\ \lim_{y \to q_{m-1}^-} f(y) & f(q_{m-1}) < \lim_{y \to q_{m-1}^-} f(y), \end{cases} \tag{6}$$

for all $x \in \mathcal{R}_m^-$ if $q_{m-1} \neq -\infty$, where $v^+ = \lim_{x \to q_{m-1}^+} f'(x)$. Figure 2 illustrates the surrogate function $f_m(x)$ with $x \in \mathcal{R}_m^+$ for the three different cases in (5).

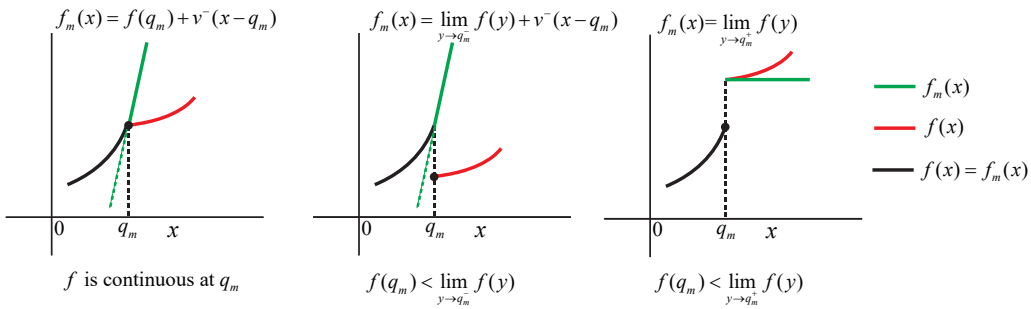


Figure 2: Illustration of three piecewise convex functions.

Example 2.1 explains how to build surrogate functions for two popular piecewise convex functions in the optimization literature. It can be observed from the definition of surrogate functions and Figure 2 that surrogate functions always have a simpler geometric structure than the original f, so it is usually the case that proximal mapping associated with the surrogate functions have closed-form expressions, given that proximal mapping associated with f has closed-form expressions. The proximal mappings associated with the surrogate functions are also given in Example 2.1.

Algorithm 1 Monotone Accelerated Proximal Gradient Descent ($\mathbf{x}^{(0)}$)

1: $\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$.
2: **for** $k = 1, \ldots,$ **do**

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k)} + \frac{t_{k-1}}{t_k}(\mathbf{z}^{(k)} - \mathbf{x}^{(k)}) + \frac{t_{k-1} - 1}{t_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}),$$
$$\mathbf{z}^{(k+1)} = \text{prox}_{sh}(\mathbf{u}^{(k)} - s\nabla g(\mathbf{u}^{(k)})),$$
$$t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2},$$
$$\mathbf{x}^{(k+1)} = \begin{cases} \mathbf{z}^{(k+1)} & \text{if } F(\mathbf{z}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k)} & \text{otherwise.} \end{cases}$$

Example 2.1 (Examples of Surrogate Functions)**.** (1) The capped-ℓ^1 penalty function $f(x) = \lambda \min\{|x|, b\}$ has three surrogate functions as $f_1 = f_3 \equiv \lambda b$, and $f_2 = \lambda |x|$. The proximal mappings for all $\{f_i\}_{i=1}^3$ has closed-form expressions, with $\text{prox}_{sf_1}(x) = \text{prox}_{sf_3}(x) = x$, and $\text{prox}_{sf_2}(x) = (1 - \lambda s / |x|)_+ x$ being the soft thresholding operator, where $(a)_+ = a$ when $a \geq 0$ and $(a)_+ = 0$ when $a < 0$. (2) The indicator penalty function $f(x) = \lambda \mathbb{I}_{\{x < \tau\}}$ has surrogate functions $f_1 \equiv \lambda$, $f_2 = f$. The proximal mappings are $\text{prox}_{sf_1}(x) = x$, $\text{prox}_{sf_2}(x) = \tau$ when $\tau - \sqrt{2\lambda s} < x < \tau$, and $\text{prox}_{sf_2}(x) = x$ otherwise except for the point $x = \tau - \sqrt{2\lambda s}$. While $\text{prox}_{sf_2}(x)$ has two values at $x = \tau - \sqrt{2\lambda s}$, it will not be evaluated at $x = \tau - \sqrt{2\lambda s}$ in our PPGD algorithm to be introduced.

Algorithm 2 Projected Proximal Gradient Descent $\left(\mathbf{x}^{(0)}\right)$

$\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$, $t_1 = 1$.
for $k = 1, \ldots,$ **do**

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k)} + \frac{t_{k-1}}{t_k}(\mathbf{z}^{(k)} - \mathbf{x}^{(k)}) + \frac{t_{k-1} - 1}{t_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}) \tag{7}$$

$$\mathbf{w}^{(k)} = \mathbf{P}_{\mathbf{x}^{(k)}}(\mathbf{u}^{(k)}) \tag{8}$$

for $i = 1, \ldots, d,$ **do**

$$\mathbf{z}_i^{(k+1)} = \mathsf{prox}_{sf_{P(\mathbf{x}_i^{(k)})}}\left(\left[\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)}\right]_i\right) \tag{9}$$

$$t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}.$$

if $F_{P(\mathbf{x}^{(k)})}(\mathbf{z}^{(k+1)}) \leq F(\mathbf{x}^{(k)})$ **then**

$$\mathbf{x}^{(k+1)} = \text{Negative-Curvature-Exploitation}\left(\mathbf{x}_i^{(k)}, \mathbf{z}_i^{(k+1)}\right) \tag{10}$$

else

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$$

Algorithm 3 Negative-Curvature-Exploitation $\left(\mathbf{x}_i^{(k)}, \mathbf{z}_i^{(k+1)}\right)$

if $P(\mathbf{z}^{(k+1)}) = P(\mathbf{x}^{(k)})$ **then**
  $\mathbf{x}^{(k+1)} = \mathbf{z}^{(k+1)}$
  return $\mathbf{x}^{(k+1)}$
Flag = false
for $i = 1, \ldots, d$ **do**
  if $P(\mathbf{z}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}) = m_i$
    if f is continuous at $q(\mathbf{w}_i^{(k)})$ and $d_{i,1} \geq w_0 d_{i,0}$ **then**
      Flag = true
    if $q(\mathbf{w}_i^{(k)})$ is a right endpoint of $\mathcal{R}_{m_i}$ and $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^-} f(y)$, or $q(\mathbf{w}_i^{(k)})$ is a left endpoint of $\mathcal{R}_{m_i}$ and $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^+} f(y)$ **then**
      Flag = true
      if $\mathcal{R}_{m_i} = \left\{q(\mathbf{w}_i^{(k)})\right\}$ **then**
        $\mathbf{x}_i^{(k+1)} = q(\mathbf{w}_i^{(k)})$.
if Flag = false **then**
  $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$

2.2 Projective Proximal Gradient Descent

We first define a basic operator P which returns the index of the convex piece that the input lies in, that is, $P\colon \mathbb{R} \to \mathbb{N}, P(x) \coloneqq m$ such that $x \in \mathcal{R}_m$. While $P(x)$ can take any of the two values when x is an endpoint at which f is continuous, this does not affect our analysis and theoretical results because $P(x) \neq P(y)$ indicates x, y are on different convex pieces . Given a vector $\mathbf{x} \in \mathbb{R}^d$, we define $P(\mathbf{x}) \in \mathbb{R}^d$ such that $[P(\mathbf{x})]_i = P(\mathbf{x}_i)$ for all $i \in [d]$. We then introduce a novel projection operator $\mathbf{P}_{\mathbf{x}}(\cdot)$ for any $\mathbf{x} \in \mathbb{R}^d$. That is, $\mathbf{P}_{\mathbf{x}}\colon \mathbb{R}^d \to \mathbb{R}^d$, and $\mathbf{P}_{\mathbf{x}}(\mathbf{u})$ for all $\mathbf{u} \in \mathbb{R}^d$ is defined as

$$[\mathbf{P}_{\mathbf{x}}(\mathbf{u})]_i \coloneqq \text{argmin}_{x \in \overline{\mathcal{R}_{P(\mathbf{x}_i)}}} |x - \mathbf{u}_i| \tag{11}$$

for all $i \in [d]$. Since $\overline{\mathcal{R}_{P(\mathbf{x}_i)}}$ is a closed set, (11) is well-defined. $\mathbf{P}_{\mathbf{x}}(\mathbf{u})$ can be computed very efficiently by setting the endpoint of $\mathcal{R}_{P(\mathbf{x}_i)}$ closer to $\mathbf{u}_i$ to $[\mathbf{P}_{\mathbf{x}}(\mathbf{u})]_i$. In fact, $[\mathbf{P}_{\mathbf{x}}(\mathbf{u})]_i$ is the projection of $\mathbf{u}_i$ onto the closed convex set $\overline{\mathcal{R}_{P(\mathbf{x}_i)}}$.

The Projected Proximal Gradient Descent (PPGD) algorithm is described in Algorithm 2. The following functions and quantities are defined which are useful for PPGD. Let $q(\mathbf{w}_i^{(k)})$ be the endpoint closest to $\mathbf{w}_i^{(k)}$ which lies in $[\mathbf{w}_i^{(k)}, \mathbf{z}_i^{(k+1)}]$ or $[\mathbf{z}_i^{(k+1)}, \mathbf{w}_i^{(k)}]$ when $P(\mathbf{z}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)})$ at iteration k of Algorithm 2. We define $F_{P(\mathbf{x})}(\mathbf{v}) := g(\mathbf{v}) + \sum_{i=1}^d f_{P(\mathbf{x}_i)}(\mathbf{v}_i)$, $d_{i,0} := \left|\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right|$, $d_{i,1} := \left|\mathbf{z}_i^{(k+1)} - q(\mathbf{w}_i^{(k)})\right|$.

It is noted that (7) is also in the regular monotone APG (Beck & Teboulle, 2009a) described in Algorithm 1 for convex problems. Compared to Algorithm 1, the projection of $\mathbf{u}^{(k)}$ onto a convex piece is used to compute $\mathbf{z}^{(k+1)}$. In addition, we have a novel Negative-Curvature-Exploitation algorithm described in Algorithm 3 which decides if PPGD should update the next iterate $\mathbf{x}^{k+1}$. In particular, the Negative-Curvature-Exploitation procedure only returns true if there is no change in convex pieces between $\mathbf{z}^{k+1}$ and $\mathbf{w}^k$ or the objective value will be decreased by an absolute positive amount by setting $\mathbf{x}^{k+1}$ to $\mathbf{z}^{k+1}$. The Negative-Curvature-Exploitation algorithm guarantees that either the next iterate $\mathbf{x}^{(k+1)}$ remains in the same convex pieces as $\mathbf{x}^{(k)}$, or the objective value is decreased by letting some coordinate of $\mathbf{x}^{(k+1)}$ transit to another convex piece. Such guarantee will be proved in our theoretical analysis in the next section.

3 Convergence Analysis

In this section, we present the analysis for the convergence rate of PPGD in Algorithm 2. Before that, we present the definition of Fréchet subdifferential which is useful for the proof of necessary lemmas.

Definition 3.1. (Fréchet subdifferential and critical points (Rockafellar & Wets, 2009)) Given a non-convex function $u\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which is a proper and lower semicontinuous function.

- For a given $\mathbf{x} \in \mathbf{dom} u$, its Fréchet subdifferential of u at $\mathbf{x}$, denoted by $\tilde{\partial} u(\mathbf{x})$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^d$ which satisfy

$$\liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \to \mathbf{x}, \mathbf{y} \in \mathbf{dom} u} \frac{u(\mathbf{y}) - u(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0.$$

- The limiting-subdifferential of u at $\mathbf{x} \in \mathbb{R}^d$, denoted by written $\partial u(\mathbf{x})$, is defined by

$$\partial u(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^d \mid \exists \mathbf{x}^k \to \mathbf{x}, u(\mathbf{x}^k) \to u(\mathbf{x}),\ \tilde{\mathbf{v}}^k \in \tilde{\partial} u(\mathbf{x}^k) \to \mathbf{v}\}.$$

The point $\mathbf{x}$ is a critical point of u if $\mathbf{0} \in \partial u(\mathbf{x})$.

The following assumption is useful for our analysis when the convex pieces have one or more endpoints at which f is continuous.

Assumption 2 (Nonvanishing Gradient at Continuous Endpoints)**.** Let q_m with $m \in [M-1]$ be an endpoint at which f is continuous. There exists a positive constant ε_0 such that the following two conditions hold. If q_m is a left endpoint, then $\inf_{\mathbf{x} \in \mathbb{R}^d, \mathbf{x}_i = q_m} |[\nabla g(\mathbf{x})]_i + p^+| \geq \varepsilon_0$ for all $i \in [d]$. If q_m is a right endpoint, then $\inf_{\mathbf{x} \in \mathbb{R}^d, \mathbf{x}_i = q_m} |[\nabla g(\mathbf{x})]_i + p^-| \geq \varepsilon_0$. Here $p^+ := \lim_{y \to q_m^+} f'(y)$, and $p^- := \lim_{y \to q_m^-} f'(y)$.

We note that Assumption 2 is rather mild. For example, when $f(x) = \lambda \min\{|x|, b\}$ is the capped-ℓ^1 penalty, then Assumption 2 holds when $\lambda > G$ and ε_0 can be set to $\lambda - G$. Such requirement for large λ is commonly used for analysis of consistency of sparse linear models, such as (Loh, 2017). In addition, when there are no endpoints at which f is continuous, then Assumption 2 is not required for the provable convergence rate of PPGD. Details are deferred to Remark 3.6.

We then have the an important Theorem 3.4, which directly follows by Lemma 3.1, Lemma 3.2, and Lemma 3.3 below. Theorem 3.4 states that the Negative-Curvature-Exploitation algorithm guarantees that, if convex pieces change across consecutive iterates $\mathbf{x}^{(k)}$ and $\mathbf{x}^{(k+1)}$, then the objective value is decreased by a positive amount.

Lemma 3.1. Suppose $s < \min\left\{\frac{J}{FG+G^2/2}, \frac{s_0}{G}\right\}$, $P(\mathbf{x}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}) = m_i$ for some $k \geq 0$ and some $i \in [d]$, and f is not continuous at $q(\mathbf{w}_i^{(k)})$. Then $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^-} f(y)$ if $q(\mathbf{w}_i^{(k)})$ is the right endpoint of $\mathcal{R}_{m_i}$ for $1 \leq m_i < M$, and $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^+} f(y)$ if $q(\mathbf{w}_i^{(k)})$ is the left endpoint of $\mathcal{R}_{m_i}$ for $1 < m_i \leq M$.

Lemma 3.2 (Decrease of f at continuous endpoint)**.** Suppose $s < \min\left\{\frac{s_0}{F+G}, \frac{\varepsilon_0}{L_g(1-w_0)(G+F)}\right\}$, $P(\mathbf{x}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}) = m_i$ for some $k \geq 0$ and some $i \in [d]$, and f is continuous at $q(\mathbf{w}_i^{(k)})$. Then $f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)})$. Furthermore, if $d_{i,1} \geq w_0 d_{i,0}$, then

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}) - \kappa_1, \tag{12}$$

where $\kappa_1 \coloneqq sCw_0\left(\varepsilon_0 - sL_g(1-w_0)(G+F)\right)$.

Lemma 3.3 (Decrease of f at discontinuous endpoints)**.** Suppose $s < \min\left\{\frac{s_0}{G+F}, \frac{J}{2F(G+F)}\right\}$, $P(\mathbf{x}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}) = m_i$ for some $k \geq 0$ and some $i \in [d]$. Suppose the following condition hold: when $q(\mathbf{w}_i^{(k)})$ is a right endpoint of $\mathcal{R}_{m_i}$, then $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^-} f(y)$; when $q(\mathbf{w}_i^{(k)})$ is a left endpoint of $\mathcal{R}_{m_i}$, then $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^+} f(y)$. Then

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}) - \kappa_2, \tag{13}$$

where $\kappa_2 \coloneqq J - sF(G+F)$.

Theorem 3.4 (Decrease of the objective function when convex pieces change)**.** Suppose

$$s < \min\left\{\frac{s_0}{F+G}, \frac{\varepsilon_0}{L_g(1-w_0)(G+F)}, \frac{J}{FG+G^2/2}, \frac{J}{2F(G+F)}\right\}. \tag{14}$$

If $P(\mathbf{x}^{(k+1)}) \neq P(\mathbf{x}^{(k)})$ for some $k \geq 0$ in the sequence $\left\{\mathbf{x}^{(k)}\right\}_{k\geq 0}$ generated by the PPGD algorithm, then

$$F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \kappa, \tag{15}$$

where $\kappa \coloneqq \min\{\kappa_1, \kappa_2\}$ with $\kappa_1 \coloneqq sCw_0\left(\varepsilon_0 - sL_g(1-w_0)(G+F)\right)$, $\kappa_2 \coloneqq J - sF(G+F)$.

Due to Theorem 3.4, there can only be a finite number of k's such that the convex pieces change across consecutive iterates $\mathbf{x}^{(k)}$ and $\mathbf{x}^{(k+1)}$. As a result, after finite iterations all the iterates generated by PPGD lie on the same convex pieces, so the nice properties of convex optimization apply to these iterates. Formally, we have the following theorem stating the convergence rate of PPGD.

Theorem 3.5 (Convergence rate of PPGD)**.** Suppose s satisfies (14), Assumption 1 and Assumption 2 hold. Then there exists a finite $k_0 \geq 1$ such that the following statements hold. (1) $P(\mathbf{x}^{(k)}) = \mathbf{m}^*$ for some $\mathbf{m}^* \in \mathbb{N}^d$ for all $k \geq k_0$. (2) Let $\mathbf{x}^*$ with $P(\mathbf{x}^*) = \mathbf{m}^*$ be a limiting point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 2, then

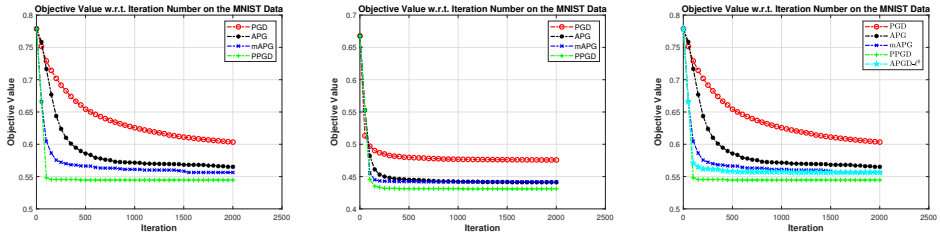
$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*) \leq \frac{4}{k^2} U^{(k_0)} \tag{16}$$

for all $k > k_0$, where $U^{(k_0)} \coloneqq \left(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\|_2^2 + t_{k_0-1}^2(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*))\right)$.

Remark 3.6. If the convex pieces $\{\mathcal{R}_m\}_{m=1}^M$ has no endpoints at which f is continuous, then Theorem 3.5 holds without requiring Assumption 2. In particular, Theorem 3.5 holds without Assumption 2 for problem (1) with capped-ℓ^1 regularizer or ℓ^0-norm regularizer.

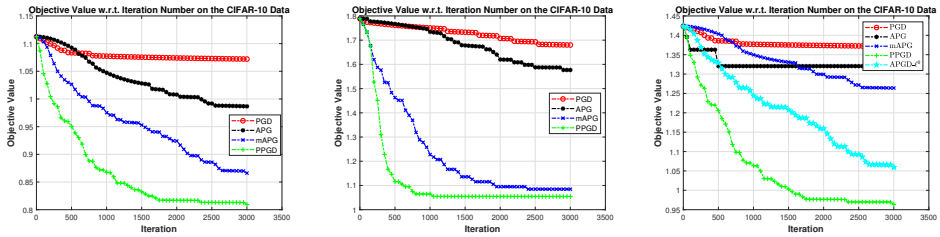
4 EXPERIMENTAL RESULTS

We conduct empirical study using the proposed PPGD methods in this section. PPGD and the baseline optimization methods are used to solve the regularized logistic regression problem, $\min_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i\mathbf{x}'x_i)) + h(\mathbf{x})$, where $g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i\mathbf{x}'x_i))$, and



(a) (b) (c)

Figure 3: (a) Illustration of the objective value with respect to the iteration number on the MNIST data. From left to right: nonnegative logistic regression, capped-ℓ regularized logistic regression, and ℓ^0-regularized logistic regression



(a) (b) (c)

Figure 4: (a) Illustration of the objective value with respect to the iteration number on the CIFAR-10 data. From left to right: nonnegative logistic regression, capped-ℓ regularized logistic regression, and ℓ^0-regularized logistic regression

$h(\mathbf{x}) = \sum_{i=1}^{d} f(\mathbf{x}_i)$ for three choices of $f(x)$: the indicator penalty, $f(x) = \lambda \mathbb{I}_{\{x<0\}}$; the capped-$\ell$ penalty, $f(x) = \lambda \min\{|x|, b\}$; the ℓ^0-norm, $f(x) = \lambda \mathbb{I}_{\{x \neq 0\}}$. Three three choices for f correspond to nonnegative logistic regression, capped-ℓ regularized logistic regression, and ℓ^0-regularized logistic regression. We conduct experiments on the MNIST handwritten digits dataset (LeCun et al., 2010) and the CIFAR-10 dataset (Krizhevsky, 2009).

The MNIST data set contains $70,000$ examples of handwritten digit images of 10 classes. Each image is of size 28×28 and represented by a 784-dimensional vector. The CIFAR-10 dataset contains 60000 images of size 32×32 with 10 classes. For each dataset, we randomly sample two classes with 5000 images in each class, which form the training data $\{\mathbf{x}_i, y_i\}_{i=1}^{10000}$ with $\{\mathbf{x}_i\}_{i=1}^{10000}$ being the sampled images and $\{y_i\}_{i=1}^{10000}$ being the corresponding class labels (± 1). We compare PPGD to Proximal Gradient Descent (PGD), the regular monotone Accelerated Proximal Gradient (APG) described in Algorithm 1, and the monotone Accelerated Proximal Gradient (mAPG) algorithm (Li & Lin, 2015). Due to the observation that monotone version of accelerated gradient methods usually converge faster than its nonmonotone counterparts, the regular nonmonotone APG and the nonmonotone Accelerated Proximal Gradient (nmAPG) method (Li & Lin, 2015) are not included in the baselines. PPGD is also compared to the fast PGD method (Yang & Yu, 2020), referred to as APGD-ℓ^0, in our experiments for ℓ^0-regularized logistic regression. Figure 3 and Figure 4 illustrate the decrease of objective value with respect to the iteration number on the MNIST dataset and the CIFAR-10 dataset respectively. It can be observed that PPGD always converges faster than all the competing baselines, evidencing the proved fast convergence results in this paper.

5 CONCLUSION

We present fast Projective Proximal Gradient Descent (PPGD) methods to solve a class of challenging nonconvex and nonsmooth optimization methods. Using a novel projection operator and a carefully designed Negative-Curvature-Exploitation algorithm, PPGD achieves Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function. The effectiveness of PPGD is evidenced by experiments.

REFERENCES

A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, Nov 2009a.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, March 2009b.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, August 2014. ISSN 0025-5610.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, March 2016a.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, Mar 2016b.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, pp. 1042–1085, 2018.

A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, Univ. Toronto, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 379–387, 2015.

Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $o(\epsilon^{-7/4})$ complexity. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12901–12916. PMLR, 17–23 Jul 2022.

Qunwei Li, Yi Zhou, Yingbin Liang, and Pramod K. Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2111–2119. PMLR, 06–11 Aug 2017.

Po-Ling Loh. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *The Annals of Statistics*, 45(2):866 – 896, 2017.

Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, Aug 2013.

Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, 2015.

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

Jianqiao Wangni and Dahua Lin. Learning sparse visual representations with leaky capped norm regularizers. *CoRR*, abs/1711.02857, 2017.

Yingzhen Yang and Jiahui Yu. Fast proximal gradient descent for a class of non-convex and non-smooth sparse learning problems. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1253–1262. PMLR, 22–25 Jul 2020.

A ALGORITHMS IN THE PAPER

A.1 PROXIMAL GRADIENT DESCENT

In the k-th iteration of PGD for $k \geq 1$, gradient descent is performed on the squared loss term $g(\mathbf{x})$ to obtain an intermediate variable as the result of gradient descent, i.e. $\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})$, where $s > 0$ is the step size, and $\frac{1}{s}$ is usually chosen to be larger than L_g. $\mathbf{x}^{(k+1)}$ is computed by proximal mapping on $\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})$.

Algorithm 4 Proximal Gradient Descent ($\mathbf{x}^{(0)}$)

1: **for** k=1,..., **do**

$$\mathbf{x}^{(k+1)} = \mathsf{prox}_{sh}\left(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})\right) \tag{17}$$

The iterations start from $k = 1$ and continue until the sequence $\{F(\mathbf{x}^{(k)})\}_k$ or $\{\mathbf{x}^{(k)}\}_k$ converges or maximum iteration number is achieved. The optimization algorithm for the ℓ^0 sparse approximation problem (1) by PGD is described in Algorithm 4. In practice, the time complexity of optimization by PGD is $\mathcal{O}(Mdn)$ where M is the number of iterations (or maximum number of iterations) for PGD.

B PROOFS

The following lemma shows that the Fermat's rule still applies to functions in the sense of Fréchet subdifferential, that is, local mininum has $\mathbf{0}$ in its Fréchet subdifferential and also its (limiting-)subdifferential.

Lemma B.1. (Fermat's rule for Fréchet subdifferential and (limiting-)subdifferential) Let u be a real-valued function defined on $\Omega \subseteq \mathbb{R}^d$, $u := \Omega \to \mathbb{R}$, has a local minimum point at $\mathbf{x} \in \Omega$. Then $\mathbf{0} \in \tilde{\partial} u(\mathbf{x}) \subseteq \partial u(\mathbf{x})$.

Proof. Since $\mathbf{x} \in \Omega$ is a local minimum point, then there exisits a neighborhood $\mathbf{x} \in \mathcal{U}(\mathbf{x})$ such that $u(\mathbf{y}) \geq u(\mathbf{x})$ for all $\mathbf{y} \in \mathcal{U}(\mathbf{x})$. It follows that

$$\liminf_{\mathbf{y}\neq\mathbf{x},\mathbf{y}\to\mathbf{x},\mathbf{y}\in\mathbf{dom}\,u} \frac{u(\mathbf{y}) - u(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \geq 0. \tag{18}$$

By (18), $\mathbf{0} \in \tilde{\partial} u(\mathbf{x})$. Because $\tilde{\partial} u(\mathbf{x}) \subseteq \partial u(\mathbf{x})$, $\mathbf{0} \in \tilde{\partial} u(\mathbf{x}) \subseteq \partial u(\mathbf{x})$. □

B.1 PROOF OF LEMMA 3.1

Proof. We prove the case when $q(\mathbf{w}_i^{(k)})$ is the right endpoint of $\mathcal{R}_{m_i}$. The case when $q(\mathbf{w}_i^{(k)})$ is the left endpoint of $\mathcal{R}_{m_i}$ can be proved in a similar manner.

Suppose the claimed result $f(q(\mathbf{w}_i^{(k)})) < \lim_{y\to q(\mathbf{w}_i^{(k)})^-} f(y)$ does not hold, so we must have $\lim_{y\to q(\mathbf{w}_i^{(k)})^+} f(y) > f(q(\mathbf{w}_i^{(k)}))$. In this case, $\mathbf{x}_i^{(k+1)} = \mathbf{z}_i^{(k+1)}$, and $\mathbf{z}_i^{(k+1)} \in (q(\mathbf{w}_i^{(k)}), \infty)$.

Define $p(v) := \frac{1}{2s}\left(v - \left[\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)}\right]_i\right)^2 + f_{m_i}(v)$. We have

$$\mathbf{z}_i^{(k+1)} = \mathsf{prox}_{sf_{P(\mathbf{x}_i^{(k)})}}\left(\left[\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)}\right]_i\right).$$

The optimality of $\mathbf{z}_i^{(k+1)}$ by Fermat's rule indicates that $\mathbf{z}_i^{(k+1)} = \left[\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)}\right]_i$, and

$$p(\mathbf{z}_i^{(k+1)}) \leq p(\mathbf{w}_i^{(k)}). \tag{19}$$

On the other hand, we have

$$p(\mathbf{w}_i^{(k)}) = \frac{s}{2}\nabla[g(\mathbf{w}^{(k)})]_i^2 + f_{m_i}(\mathbf{w}_i^{(k)}), \quad p(\mathbf{z}_i^{(k+1)}) \geq f_{m_i}(\mathbf{z}_i^{(k+1)}) = \lim_{y\to q(\mathbf{w}_i^{(k)})^+} f(y), \tag{20}$$

and

$$\left|f_{m_i}(\mathbf{w}_i^{(k)}) - f_{m_i}(q(\mathbf{w}_i^{(k)}))\right| = \left|f(\mathbf{w}_i^{(k)}) - f(q(\mathbf{w}_i^{(k)}))\right| \leq F\left|\mathbf{w}_i^{(k)} - q(\mathbf{w}_i^{(k)})\right|$$
$$\leq F\left|\mathbf{w}_i^{(k)} - \mathbf{z}_i^{(k+1)}\right| \leq sF[\nabla g(\mathbf{w}^{(k)})]_i \leq sFG. \tag{21}$$

(21) follows because $\mathbf{w}_i^{(k)} \in (q(\mathbf{w}_i^{(k)}) - s_0, q(\mathbf{w}_i^{(k)}))$ due to $s < \frac{s_0}{G}$.

It follows by (20) and (21) that

$$p(\mathbf{z}_i^{(k+1)}) \geq f(q(\mathbf{w}_i^{(k)})) + J > f(q(\mathbf{w}_i^{(k)})) + sFG + \frac{s}{2}G^2 \geq h(\mathbf{w}_i^{(k)}) \tag{22}$$

because $s < \frac{J}{FG+G^2/2}$. This contradiction shows that we must have $\lim_{y\to q(\mathbf{w}_i^{(k)})^+} f(y) \leq f(q(\mathbf{w}_i^{(k)}))$. □

B.2 Proof of Lemma 3.2

Proof. Let $q = q(\mathbf{w}_i^{(k)})$. We must have $\mathbf{x}_i^{(k+1)} = \mathbf{z}_i^{(k+1)}$ with the PPGD algorithm and the Negative-Curvature-Exploitation algorithm described in Algorithm 2 and Algorithm 3. Without loss of generality, we assume $\mathbf{x}_i^{(k+1)} \in \mathcal{R}_{m_i}^+$. The case that $\mathbf{x}_i^{(k+1)} \in \mathcal{R}_{m_i}^-$ can be proved in a similar manner.

We have

$$f_{m_i}(\mathbf{x}_i^{(k+1)}) = f(q) + v^-\left(\mathbf{x}_i^{(k+1)} - q\right) \tag{23}$$

by the definition (5), where $v^- = \lim_{x\to q^-} f'(x)$. On the other hand, we have

$$\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)} = -s\left([\nabla g(\mathbf{w}^{(k)})]_i + f'_{m_i}(\mathbf{z}_i^{(k+1)})\right) = s\left([\nabla g(\mathbf{w}^{(k)})]_i + v^-\right). \tag{24}$$

It follows by (24) that

$$\left|\mathbf{x}_i^{(k+1)} - q\right| = \left|\mathbf{z}_i^{(k+1)} - q\right| \leq \left|\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right| = s\left|[\nabla g(\mathbf{w}^{(k)})]_i + v^-\right| \leq s(G+F). \tag{25}$$

As a result, $s < \frac{s_0}{G+F}$ guarantees that $\mathbf{x}_i^{(k+1)} \in (q, q+s_0)$. Therefore, we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f(x) + f'(\mathbf{x}_i^{(k+1)})\left(\mathbf{x}_i^{(k+1)} - x\right) \tag{26}$$

when $x \in (q, \mathbf{x}_i^{(k+1)})$. Letting $x \to q$ in (26), we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f(q) + v^+\left(\mathbf{x}_i^{(k+1)} - q\right), \tag{27}$$

where $v^+ = \lim_{x\to q^+} f'(x)$. By the negative curvature condition in Assumption 1(d), we have $v^- - v^+ \geq C > 0$, therefore,

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{x}_i^{(k+1)}) - C\left(\mathbf{x}_i^{(k+1)} - q\right)$$
$$\leq f_{m_i}(\mathbf{x}_i^{(k+1)}) - Cw_0\left(\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right). \tag{28}$$

Because $\mathbf{z}_i^{(k+1)} \in \mathcal{R}_{m_i}^+$, $\mathbf{z}_i^{(k+1)} \geq \mathbf{w}_i^{(k)}$. By (28) we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{x}_i^{(k+1)}) = f_{m_i}(\mathbf{z}_i^{(k+1)}). \tag{29}$$

Let $\tilde{\mathbf{w}} \in \mathbb{R}^d$ with $\tilde{\mathbf{w}}_j = \mathbf{w}_j^{(k)}$ for all $j \neq i$, and $\tilde{\mathbf{w}}_i = q$. By the definition of $\tilde{\mathbf{w}}$ and (24), we have

$$\left\|\mathbf{w}^{(k)} - \tilde{\mathbf{w}}\right\|_2 = \left|\mathbf{w}_i^{(k)} - q\right| \leq (1-w_0)\left|\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right|$$
$$= s(1-w_0)\left|[\nabla g(\mathbf{w}^{(k)})]_i + v^-\right| \leq s(1-w_0)(G+F). \tag{30}$$

We have

$$\begin{aligned}\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)} &= \left|\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right| = s\left|[\nabla g(\mathbf{w}^{(k)})]_i + v^-\right| \\ &= s\left|[\nabla g(\mathbf{w}^{(k)}) - \nabla g(\tilde{\mathbf{w}})]_i + [\nabla g(\tilde{\mathbf{w}})]_i + v^-\right| \\ &\geq s\left(\varepsilon_0 - L_g\left\|\mathbf{w}^{(k)} - \tilde{\mathbf{w}}\right\|_2\right) \\ &\geq s\left(\varepsilon_0 - sL_g(1-w_0)(G+F)\right), \end{aligned} \tag{31}$$

where the last inequality follows from (30).

It follows by (28) and (31) that

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{x}_i^{(k+1)}) - sCw_0\left(\varepsilon_0 - sL_g(1-w_0)(G+F)\right). \tag{32}$$

Noting that $\mathbf{x}_i^{(k+1)} = \mathbf{z}_i^{(k+1)}$ completes the proof. □

B.3 Proof of Lemma 3.3

Proof. Let $q = q(\mathbf{w}_i^{(k)})$, and we have $\mathbf{x}_i^{(k+1)} = \mathbf{z}_i^{(k+1)}$. Without loss of generality, we assume $\mathbf{x}_i^{(k+1)} \in \mathcal{R}_{m_i}^+$. The case that $\mathbf{x}_i^{(k+1)} \in \mathcal{R}_{m_i}^-$ can be proved in a similar manner.

We first consider the case that $\mathcal{R}_{m_i}$ is not a single-point set, that is, $\mathcal{R}_{m_i} \neq \{q\}$.

According to the definition of surrogate function (5), we have

$$f_{m_i}(\mathbf{x}_i^{(k+1)}) = \lim_{y \to q^-} f(y) + v^-(\mathbf{x}_i^{(k+1)} - q), \tag{33}$$

where $v^- = \lim_{x\to q^-} f'(x)$. If $\mathbf{x}_i^{(k+1)} \neq q$, applying the argument in (25) in the proof of Lemma 3.2, $s < \frac{s_0}{G+F}$ guarantees that $\mathbf{x}_i^{(k+1)} \in (q, q+s_0)$. So we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f(x) + f'(\mathbf{x}_i^{(k+1)})\left(\mathbf{x}_i^{(k+1)} - x\right) \tag{34}$$

when $x \in (q, \mathbf{x}_i^{(k+1)})$. Letting $x \to q$ in (34), we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f(q) + v^+\left(\mathbf{x}_i^{(k+1)} - q\right), \tag{35}$$

where $v^+ = \lim_{x\to q^+} f'(x)$. In addition,

It follows by (35) and (25) in the proof of Lemma 3.2 that

$$f(\mathbf{x}_i^{(k+1)}) \leq f(q) + sF(G+F). \tag{36}$$

We note that (36) holds for all $\mathbf{x}_i^{(k+1)} \in [q, q+s_0)$ Moreover, it follows by (33) and (25) in the proof of Lemma 3.2 that

$$f_{m_i}(\mathbf{x}_i^{(k+1)}) \geq \lim_{y\to q^-} f(y) - \left|v^-(\mathbf{x}_i^{(k+1)} - q)\right| \geq \lim_{y\to q^-} f(y) - sF(G+F) \geq f(q) + J - sF(G+F). \tag{37}$$

Combining (36) and (37), we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{x}_i^{(k+1)}) - (J - 2sF(G+F)) = f_{m_i}(\mathbf{z}_i^{(k+1)}) - (J - 2sF(G+F)). \tag{38}$$

If $\mathcal{R}_{m_i} = \{q\}$, then $\mathbf{x}_i^{(k+1)} = q$ and $f(\mathbf{x}_i^{(k+1)}) = f(q)$. By the fact that

$$f_{m_i}(\mathbf{z}_i^{(k+1)}) = \lim_{y\to q^-} f(y) + v^-(\mathbf{z}_i^{(k+1)} - q)$$

and $\left|\mathbf{z}_i^{(k+1)} - q\right| \leq \left|\mathbf{z}_i^{(k+1)} - \mathbf{w}_i^{(k)}\right|$, following the argument similar to (37) we have

$$f_{m_i}(\mathbf{z}_i^{(k+1)}) \geq f(q) + J - sF(G+F) = f(\mathbf{x}_i^{(k+1)}) + J - sF(G+F).$$

It follows that

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}) - (J - sF(G+F)). \tag{39}$$

The proof is completed by combining (38) and (39).

□

B.4 PROOF OF THEOREM 3.4

Proof. We split $[d]$ into three disjoint subsets, $[d] = S_1 \cup S_2 \cup S_3$, and the three subsets are defined by

$$S_1 \coloneqq \left\{ i \in [d] \colon P(\mathbf{x}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}), f \text{ is continuous at } q(\mathbf{w}_i^{(k)}) \right\}$$

$$S_2 \coloneqq \left\{ i \in [d] \colon P(\mathbf{x}_i^{(k+1)}) \neq P(\mathbf{x}_i^{(k)}), f \text{ is not continuous at } q(\mathbf{w}_i^{(k)}) \right\}, \tag{40}$$

$$S_3 \coloneqq \left\{ i \in [d] \colon P(\mathbf{x}_i^{(k+1)}) = P(\mathbf{x}_i^{(k)}) \right\}. \tag{41}$$

Let $m_i = P(\mathbf{x}_i^{(k)})$. According to Lemma 3.2, for all $i \in S_1$ such that $d_{i,1} \geq w_0 d_{i,0}$, we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}) - \kappa_1. \tag{42}$$

In addition, for all $i \in S_1$, we have

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}). \tag{43}$$

According to Lemma 3.1, S_2 satisfies the condition that for all $i \in S_2$, when $q(\mathbf{w}_i^{(k)})$ is a right endpoint of $\mathcal{R}_{m_i}$, then $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^-} f(y)$; when $q(\mathbf{w}_i^{(k)})$ is a left endpoint of $\mathcal{R}_{m_i}$, then $f(q(\mathbf{w}_i^{(k)})) < \lim_{y \to q(\mathbf{w}_i^{(k)})^+} f(y)$. By Lemma 3.3,

$$f(\mathbf{x}_i^{(k+1)}) \leq f_{m_i}(\mathbf{z}_i^{(k+1)}) - \kappa_2. \tag{44}$$

for all $i \in S_2$.

For all $i \in S_3$, we have

$$f(\mathbf{x}_i^{(k+1)}) = f_{m_i}(\mathbf{x}_i^{(k+1)}) = f_{m_i}(\mathbf{z}_i^{(k+1)}). \tag{45}$$

The Negative-Curvature-Exploitation algorithm described in Algorithm 3 guarantees that when $P(\mathbf{x}^{(k+1)}) \neq P(\mathbf{x}^{(k)})$, there exists at lease one $i \in S_1 \cup S_2$ such that (42) or (44) holds. It follows by (42)-(45) and the above argument that

$$\sum_{i=1}^{d} f(\mathbf{x}_i^{(k+1)}) \leq \sum_{i=1}^{d} f_{m_i}(\mathbf{z}_i^{(k+1)}) - \kappa = \sum_{i=1}^{d} f_{m_i}(\mathbf{x}_i^{(k+1)}) - \kappa, \tag{46}$$

where the last equality follows from $\mathbf{x}^{(k+1)} = \mathbf{z}^{(k+1)}$. It follows from (46) that

$$F(\mathbf{x}_i^{(k+1)}) \leq F_{P(\mathbf{x}^{(k)})}(\mathbf{x}_i^{(k+1)}) - \kappa.$$

The PPGD algorithm guarantees that

$$F_{P(\mathbf{x}^{(k)})}(\mathbf{z}^{(k+1)}) = F_{P(\mathbf{x}^{(k)})}(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}).$$

Therefore, $F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \kappa$ which completes the proof. □

B.5 PROOF OF THEOREM 3.5

Proof. According to Theorem 3.4, when $P(\mathbf{x}^{(k+1)}) \neq P(\mathbf{x}^{(k)})$, $F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \kappa$. Because $\inf_{x \in \mathbb{R}^d} F(x) > -\infty$, we can only have finite number of k's such that $F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \kappa$. As a result, there must exists a finite $k_0 \in \mathbb{N}$ such that $P(\mathbf{x}^{(k)}) = \mathbf{m}^* \in \mathbb{N}^d$ for all $k \geq k_0$. It follows that $\mathbf{x}_i^* \in \overline{\mathcal{R}_{\mathbf{m}_i^*}}$.

Define $h_{\mathbf{m}}(\mathbf{x}) \coloneqq \sum_{i=1}^{d} f_{\mathbf{m}_i}(\mathbf{x}_i)$ for $\mathbf{m} \in \mathbb{N}^d$ and $\mathbf{m}_i \in [M]$ for all $i \in [d]$, and $F_{\mathbf{m}} \coloneqq g + h_{\mathbf{m}}$. Note that for all $i \in [d]$, $f_{\mathbf{m}_i^*}$ is convex except for the third case in (5) or (6). In such a case, either event ①: $f_{\mathbf{m}_i^*}(x) = \lim_{y \to q_{\mathbf{m}_i^*}^+} f(y)$ for $x > q_{\mathbf{m}_i^*}$ and $f(q_{\mathbf{m}_i^*}) < \lim_{y \to q_{\mathbf{m}_i^*}^+} f(y)$, or event ②: $f_{\mathbf{m}_i^*}(x) = \lim_{y \to q_{\mathbf{m}_i^*-1}^-} f(y)$ for $x < q_{\mathbf{m}_i^*-1}$ and $f(q_{\mathbf{m}_i^*-1}) < \lim_{y \to q_{\mathbf{m}_i^*-1}^-} f(y)$. It

follows by the proof of Lemma 3.1 that all the sequences $\left\{\mathbf{x}_i^{(k)}\right\}_{k\geq k_0}$ and $\left\{\mathbf{z}_i^{(k)}\right\}_{k\geq k_0+1}$ satisfies $\left\{\mathbf{x}_i^{(k)}\right\}_{k\geq k_0} \subseteq (-\infty, q_{\mathbf{m}_i^*}]$ and $\left\{\mathbf{z}_i^{(k)}\right\}_{k\geq k_0+1} \subseteq (-\infty, q_{\mathbf{m}_i^*}]$ if event ① happens, and $\left\{\mathbf{x}_i^{(k)}\right\}_{k\geq k_0} \subseteq [q_{\mathbf{m}_i^*-1}, +\infty)$ and $\left\{\mathbf{z}_i^{(k)}\right\}_{k\geq k_0+1} \subseteq [q_{\mathbf{m}_i^*-1}, +\infty)$ if event ② happens. For all $i \in [d]$, define $\mathcal{R}_i$ as the region over which $f_{\mathbf{m}_i^*}$ is convex. It is clear that $\mathcal{R}_i = \mathbb{R}$ if event ① and ② do not happen for $f_{\mathbf{m}_i^*}$. If only event ① happens for $f_{\mathbf{m}_i^*}$, then $\mathcal{R}_i = (-\infty, q_{\mathbf{m}_i^*}]$. If only event ② happens, $\mathcal{R}_i = [q_{\mathbf{m}_i^*-1}, +\infty)$. If both event ① and ② happen, then $\mathcal{R}_i = [q_{\mathbf{m}_i^*-1}, q_{\mathbf{m}_i^*}]$. As a result, we have

$$f_{\mathbf{m}_i^*}(\mathbf{z}_i^{(k+1)}) \leq f(v) + p(\mathbf{z}_i^{(k+1)} - v)$$

for all $i \in [d]$, $v \in \mathcal{R}_i$, and all $p \in \tilde{\partial} f(\mathbf{z}_i^{(k+1)})$. It follows that if $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathcal{R}_i$ for all $i \in [d]$, then

$$h_{\mathbf{m}_i^*}(\mathbf{z}^{(k+1)}) \leq h(\mathbf{v}) + \mathbf{p}(\mathbf{z}^{(k+1)} - \mathbf{v}) \tag{47}$$

for all $\mathbf{p} \in \tilde{\partial} h_{\mathbf{m}_i^*}$.

Because $\mathbf{z}_i^{(k+1)} = \mathsf{prox}_{sf_{P(\mathbf{x}_i^{(k)})}}\left(\left[\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)}\right]_i\right)$ in (9) of Algorithm 2, it follows by the optimality of $\mathbf{z}_i^{(k+1)}$ that

$$\frac{1}{s}(\mathbf{w}^{(k)} - \mathbf{z}^{k+1}) - \nabla g(\mathbf{w}^k) \in \tilde{\partial} h(\mathbf{z}^{(k+1)}). \tag{48}$$

It follows by (47) and (48) that

$$h_{\mathbf{m}_i^*}(\mathbf{z}^{(k+1)}) \leq h(\mathbf{v}) + \left(\frac{1}{s}(\mathbf{w}^{(k)} - \mathbf{z}^{k+1}) - \nabla g(\mathbf{w}^k)\right)(\mathbf{z}^{(k+1)} - \mathbf{v}) \tag{49}$$

for any $\mathbf{v} \in \mathbb{R}^d$ such that $\mathbf{v}_i \in \mathcal{R}_i$ for all $i \in [d]$. For such $\mathbf{v}$, we have

$$\begin{aligned}
F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) &\leq g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{z}^{(k+1)} - \mathbf{v}\rangle + \frac{L_g}{2}\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) \\
&\overset{①}{\leq} g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{z}^{(k+1)} - \mathbf{v}\rangle + \frac{L_g}{2}\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2 + h_{\mathbf{m}^*}(\mathbf{v}) \\
&\quad + \langle \nabla g(\mathbf{w}^{(k)}) + \frac{1}{s}(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), \mathbf{v} - \mathbf{z}^{(k+1)}\rangle \\
&= F_{\mathbf{m}^*}(\mathbf{v}) + \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{z}^{(k+1)}\rangle + \frac{L_g}{2}\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2 \\
&\leq F_{\mathbf{m}^*}(\mathbf{v}) + \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)}\rangle - \frac{1}{s}\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + \frac{L_g}{2}\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2 \\
&= F_{\mathbf{m}^*}(\mathbf{v}) + \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)}\rangle - \left(\frac{1}{s} - \frac{L_g}{2}\right)\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2.
\end{aligned} \tag{50}$$

Here ① follows from (49).

Let $\mathbf{v} = \mathbf{x}^{(k)}$ and $\mathbf{v} = \mathbf{x}^*$ in (50), we have

$$F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) \leq F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) + \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{x}^{(k)} - \mathbf{w}^{(k)}\rangle - \left(\frac{1}{s} - \frac{L}{2}\right)\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2, \tag{51}$$

and

$$F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) \leq F_{\mathbf{m}^*}(\mathbf{x}^*) + \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{x}^* - \mathbf{w}^{(k)}\rangle - \left(\frac{1}{s} - \frac{L}{2}\right)\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2. \tag{52}$$

(51)$\times(t_k - 1)+$ (52), we have

$$\begin{aligned}
&t_k F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) - (t_k - 1)F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) - F_{\mathbf{m}^*}(\mathbf{x}^*) \\
&\leq \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)} - t_k\left(\frac{1}{s} - \frac{L}{2}\right)\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2.
\end{aligned} \tag{53}$$

It follows that

$$t_k\big(F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) - (t_k-1)\big(F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big)$$
$$\leq \frac{1}{s}\langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, (t_k-1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)}\rangle - t_k\left(\frac{1}{s} - \frac{L}{2}\right)\left\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2. \tag{54}$$

Multiplying both sides of (54) by t_k, since $t_k^2 - t_k = t_{k-1}^2$, we have

$$t_k^2\left(F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\right) - t_{k-1}^2\left(F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\right)$$
$$\leq \frac{1}{s}\langle t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), (t_k-1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)}\rangle - (\frac{1}{s} - \frac{L}{2})\|t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)})\|_2^2$$
$$\leq \frac{1}{s}\langle t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), (t_k-1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)}\rangle - \frac{1}{2s}\left\|t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)})\right\|_2^2$$
$$= \frac{1}{2s}\left(\left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{w}^{(k)} + \mathbf{x}^*\right\|_2^2 - \left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{z}^{(k+1)} + \mathbf{x}^*\right\|_2^2\right). \tag{55}$$

Note that $\mathbf{w}^{(k)} = \mathbf{P}_{\mathbf{x}^{(k)}}(\mathbf{u}^{(k)})$ by (8) in Algorithm 2. Because $t_k > 1$ for all $k > 1$, $(1-1/t_k)\mathbf{x}^{(k)} + 1/t_k\mathbf{x}^*$ is a convex combination of $\mathbf{x}^{(k)}$ and $\mathbf{x}^*$ with $P(\mathbf{x}^{(k)}) = P(\mathbf{x}^*) = \mathbf{m}^*$, it follows that $P((1-1/t_k)\mathbf{x}^{(k)} + 1/t_k\mathbf{x}^*) = \mathbf{m}^*$

We have

$$\begin{aligned}\left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{w}^{(k)} + \mathbf{x}^*\right\|_2 &= t_k\left\|\left(1 - \frac{1}{t_k}\right)\mathbf{x}^{(k)} + \frac{1}{t_k}\mathbf{x}^* - \mathbf{w}\right\|_2 \\ &= t_k\left\|\mathbf{P}_{\mathbf{x}^{(k)}}\left(\left(1 - \frac{1}{t_k}\right)\mathbf{x}^{(k)} + \frac{1}{t_k}\mathbf{x}^*\right) - \mathbf{P}_{\mathbf{x}^{(k)}}(\mathbf{u}^{(k)})\right\|_2 \\ &\leq t_k\left\|\left(1 - \frac{1}{t_k}\right)\mathbf{x}^{(k)} + \frac{1}{t_k}\mathbf{x}^* - \mathbf{u}^{(k)}\right\|_2 \\ &\leq \left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{u}^{(k)} + \mathbf{x}^*\right\|_2,\end{aligned} \tag{56}$$

where the second last inequality follows from the contraction property of projection onto closed convex sets in Hilbert spaces.

It follows by (55) and (56), that

$$t_k^2\big(F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) - t_{k-1}^2\left(F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\right)$$
$$\leq \frac{1}{2s}\left(\left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{u}^{(k)} + \mathbf{x}^*\right\|_2^2 - \left\|(t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{z}^{(k+1)} + \mathbf{x}^*\right\|_2^2\right). \tag{57}$$

Define $\mathbf{Q}^{(k+1)} = (t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{z}^{(k+1)} + \mathbf{x}^*$, then $\mathbf{Q}^{(k)} = (t_{k-1}-1)\mathbf{x}^{(k-1)} - t_{k-1}\mathbf{z}^{(k)} + \mathbf{x}^*$. It can be verified that $\mathbf{Q}^{(k)} = (t_k-1)\mathbf{x}^{(k)} - t_k\mathbf{u}^{(k)} + \mathbf{x}^*$. Therefore,

$$t_k^2\big(F_{\mathbf{m}^*}(\mathbf{z}^{(k+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) - t_{k-1}^2\big(F_{\mathbf{m}^*}(\mathbf{x}^{(k)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) \leq \frac{1}{2s}\left(\left\|\mathbf{Q}^{(k)}\|_2^2 - \|\mathbf{Q}^{(k+1)}\right\|_2^2\right). \tag{58}$$

Summing (58) over $k = k_0, \ldots, m$ for $m \geq k_0$, we have

$$\begin{aligned}t_m^2\big(F_{\mathbf{m}^*}(\mathbf{z}^{(m+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) - t_{k_0-1}^2\big(F_{\mathbf{m}^*}(\mathbf{x}^{(k_0)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\big) &\leq \frac{1}{2s}\big(\|\mathbf{Q}^{(k_0)}\|_2^2 - \|\mathbf{Q}^{(m+1)}\|_2^2\big) \leq \frac{1}{2s}\left\|\mathbf{Q}^{(k_0)}\right\|_2^2 \\ &= \frac{1}{2s}\left\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\right\|_2^2.\end{aligned} \tag{59}$$

Since $t_k \geq \frac{k+1}{2}$ for $k \geq 1$, it follows from (59) that

$$\begin{aligned}F_{\mathbf{m}^*}(\mathbf{z}^{(m+1)}) - F_{\mathbf{m}^*}(\mathbf{x}^*) &\leq \frac{4}{(m+1)^2}\left(\frac{1}{2s}\left\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\right\|_2^2 + t_{k_0-1}^2\left(F_{\mathbf{m}^*}(\mathbf{x}^{(k_0)}) - F_{\mathbf{m}^*}(\mathbf{x}^*)\right)\right) \\ &\triangleq \frac{4}{(m+1)^2}U^{(k_0)}.\end{aligned} \tag{60}$$

Noting that $F(\mathbf{x}^{(m+1)}) = F_{\mathbf{m}^*}(\mathbf{x}^{(m+1)}) \leq F_{\mathbf{m}^*}(\mathbf{z}^{(m+1)})$, $F(\mathbf{x}^{(k_0)}) = F_{\mathbf{m}^*}(\mathbf{x}^{(k_0)})$, and $F(\mathbf{x}^*) = F_{\mathbf{m}^*}(\mathbf{x}^*)$, we have $F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \leq \frac{4}{(m+1)^2} U^{(k_0)}$. Replacing $m+1$ with k completes completes the proof.

□