Supplementary Document for Fast Proximal Gradient Descent for A Class of Non-convex and Non-smooth Sparse Learning Problems

Yingzhen Yang
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
yingzhen.yang@asu.edu

Jiahui Yu
Beckman Institute
University of Illinois at Urbana-Champaign
jyu79@illinois.edu

1 ALGORITHMS IN THE PAPER

1.1 Proximal Gradient Descent

The optimization problem studied in this paper is

\[ \min_{x \in \mathbb{R}^n} F(x) = g(x) + h(x), \]

where \( h(x) \triangleq \lambda \|x\|_0, \lambda > 0 \) is a weighting parameter.

\[
x^{(k+1)} = \text{prox}_{sh}(x^{(k)} - s\nabla g(x^{(k)}))
\]

\[
= \arg \min_{v \in \mathbb{R}^n} \frac{1}{2s} \|v - (x^{(k)} - s\nabla g(x^{(k)}))\|^2 + \lambda \|v\|_0
\]

\[
= T_{\frac{1}{2s}}(x^{(k)} - s\nabla g(x^{(k)})),
\]

(2)

Algorithm 1 Proximal Gradient Descent for the \( \ell^0 \) Regularization Problem (1)

\textbf{Input:} The weighting parameter \( \lambda \), the initialization \( x^{(0)} \).

\begin{enumerate}
  \item for \( k = 0, \ldots, \) do
  \item Update \( x^{(k+1)} \) according to (2)
  \item end for
\end{enumerate}

\textbf{Output:} Obtain the sparse solution \( \hat{x} \) upon the termination of the iterations.

1.2 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection

\[
u^{(k)} = x^{(k)} + \frac{t_{k-1}}{t_k}(x^{(k)} - x^{(k-1)}),
\]

(3)

\[
w^{(k)} = P_{\text{supp}(x^{(k)})}(u^{(k)}),
\]

(4)

\[
x^{(k+1)} = \text{prox}_{sh}(w^{(k)} - s\nabla g(w^{(k)})),
\]

(5)

\[
t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2},
\]

(6)

Algorithm 2 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection for the \( \ell^0 \) Regularization Problem (1)

\textbf{Input:}

The weighting parameter \( \lambda \), the initialization \( x^{(0)} \),
\( z^{(0)} = x^{(0)}, t_0 = 0. \)

\begin{enumerate}
  \item for \( k = 1, \ldots, \) do
  \item Update \( u^{(k)}, w^{(k)}, x^{(k+1)}, t_{k+1} \) according to (3), (4), (5), (6) respectively.
  \item end for
\end{enumerate}

\textbf{Output:} Obtain the sparse solution \( \hat{x} \) upon the termination of the iterations.

1.3 Monotone Accelerated Proximal Gradient Descent with Support Projection

\[
u^{(k)} = x^{(k)} + \frac{t_{k-1}}{t_k}(z^{(k)} - x^{(k)}),
\]

(7)

\[
w^{(k)} = P_{\text{supp}(z^{(k)})}(u^{(k)}),
\]

(8)

\[
z^{(k+1)} = \text{prox}_{sh}(w^{(k)} - s\nabla g(w^{(k)})),
\]

(9)

\[
t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2},
\]

(10)

\[
x^{(k+1)} = \begin{cases} z^{(k+1)} & \text{if } F(z^{(k+1)}) \leq F(x^{(k)}) \\
x^{(k)} & \text{otherwise} \end{cases}
\]

(11)

2 PROOFS

\textbf{Lemma 1.} (Support shrinkage for proximal gradient descent in Algorithm 1 and sufficient decrease of the objective function) If \( s \leq \min \left( \frac{2\lambda}{L}, \frac{1}{L} \right) \), then

\[
\text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)}), \quad k \geq 0,
\]

(12)

namely the support of the sequence \( \{x^{(k)}\}_k \) shrinks.
Moreover, the sequence of the objective \( \{F(x^{(k)})\}_k \) is
**Algorithm 3** Monotone Accelerated Proximal Gradient Descent with Support Projection for the $f^0$ Regularization Problem (1)

**Input:**
- The weighting parameter $\lambda$, the initialization $x^{(0)}$, $z^{(1)} = x^{(1)} = x^{(0)}$, $t_0 = 0$.

1: **for** $k = 1, \ldots$ **do**
2: \hspace{1em} Update $u^{(k)}$, $w^{(k)}$, $z^{(k+1)}$, $t_{k+1}$, $x^{(k+1)}$ according to (7), (8), (9), (10), and (11) respectively.
3: **end for**

**Output:** Obtain the sparse solution $\hat{x}$ upon the termination of the iterations.

nonincreasing, and the following inequality holds for $k \geq 0$:

$$F(x^{(k+1)}) \leq F(x^{(k)}) - \left(\frac{1}{2s} - \frac{L}{2}\right)\|x^{(k+1)} - x^{(k)}\|_2^2.$$  \hfill (13)

**Proof of Lemma 1.** We prove this Lemma by mathematical induction.

With $k \geq 0$, we first show that $\text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)})$, i.e. the support of the sequence shrinks. To see this, let $\hat{x}^{(k+1)} = x^{(k)} - s \nabla g(x^{(k)})$. Since $\|y - Dx^{(k)}\|_2^2 = x_0$, let $q^{(k)} = -s \nabla g(x^{(k)}) = -2s(D^T Dx^{(k)} - D^T y)$, then

$$|x_j^{(k+1)}| \leq |q_j^{(k)}|_\infty \leq sG,$$

where $j$ is the index for any zero element of $x^{(k)}$, namely $1 \leq j \leq d, j \notin \text{supp}(x^{(k)})$. Now $|x_j^{(k+1)}| < \sqrt{2} \lambda s$, and it follows that $x_j^{(k+1)} = 0$ due to the update rule (2). Therefore, the zero elements of $x^{(k)}$ remain unchanged in $x^{(k+1)}$, and $\text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)})$ for $k \geq 0$.

Since

$$x^{(k+1)} = \arg\min_{v \in R^d} \frac{1}{2s} \|v - \hat{x}^{(k+1)}\|_2^2 + h(v),$$

let $v = x^{(k)}$, we have

$$\frac{1}{2s} \|x^{(k+1)} - \hat{x}^{(k+1)}\|_2^2 + h(x^{(k+1)})$$

$$\leq \frac{1}{2s} \|s \nabla g(x^{(k)})\|_2^2 + h(x^{(k)}),$$

which is equivalent to

$$\langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{1}{2s} \|x^{(k+1)} - x^{(k)}\|_2^2 + h(x^{(k+1)})$$

$$\leq h(x^{(k)}).$$  \hfill (14)

In addition, since $L$ is the Lipschitz constant for $\nabla g$,

$$g(x^{(k+1)}) \leq g(x^{(k)}) + \langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2s} \|x^{(k+1)} - x^{(k)}\|_2^2.$$

Combining (15) and (16), we have

$$g(x^{(k+1)}) + h(x^{(k+1)}) \leq g(x^{(k)}) + h(x^{(k)})$$

$$- \left(\frac{1}{2s} - \frac{L}{2}\right)\|x^{(k+1)} - x^{(k)}\|_2^2.$$  \hfill (17)

Now (12) and (13) hold for $k \geq 0$. Since the sequence $\{F(x^{(k)})\}_k$ is decreasing with lower bound 0, it must converge. \hfill \Box

**Lemma A.** (Lemma 1 in Laurent and Massart (2000)) Let $Y_1, Y_2, \ldots Y_D$ be i.i.d. Gaussian random variables with 0 mean and unit variance, and $a_1, a_2, \ldots a_D$ be $D$ positive numbers. Define $Z = \sum_{i=1}^D a_i(Y_i^2 - 1)$ and $a = [a_1, a_2, \ldots a_D]^T$, then for any $t > 0$,

$$\Pr[Z \geq 2\|a\|_2 \sqrt{t} + 2\|a\|_\infty t] \leq e^{-t}. \hfill (18)$$

**Lemma B.** (Spectrum bound for Gaussian random matrix, Theorem I.I.13 in Davidson and Szarek (2001)) Suppose $A \in R^{m \times n}$ ($m \geq n$) is a random matrix whose entries are i.i.d. samples generated from the standard Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. Then

$$1 - \sqrt{\frac{n}{m}} \leq \mathbb{E}[\sigma_n(A)] \leq \mathbb{E}[\sigma_1(A)] \leq 1 + \sqrt{\frac{n}{m}}. \hfill (19)$$

Also, for any $t > 0$,

$$\Pr[\sigma_n(A) \leq 1 - \sqrt{\frac{n}{m}} - t] < e^{-\frac{nt^2}{2}},$$

$$\Pr[\sigma_1(A) \geq 1 + \sqrt{\frac{n}{m}} + t] < e^{-\frac{nt^2}{2}}. \hfill (20)$$

**Theorem 1.** Suppose $D \in R^{d \times n}$ ($n \geq d$) is a random matrix whose elements are i.i.d. samples from the standard Gaussian distribution $\mathcal{N}(0, 1)$. Then with probability at least $1 - e^{-\frac{nt^2}{2}} - ne^{-t}$,

$$2\lambda \geq \frac{1}{L} \hfill (21)$$

if

$$n \geq \left(\sqrt{d} + t + \sqrt{(d + 2\sqrt{dt} + 2t)(x_0 + \lambda|S|)}\right)^2, \hfill (22)$$

and $t$ can be chosen as $t_0 \log n$ for $t_0 > 0$ to ensure that (22) holds and (21) holds with high probability.

**Proof of Theorem 1.** According to Lemma B, for any $t > 0$, with probability at least $1 - e^{-\frac{nt^2}{2}}$,

$$\sigma_{\max}(D) > \sqrt{n} - \sqrt{d} - t. \hfill (23)$$
Also, by Lemma A, for any $1 \leq i \leq n$ and $t > 0$, with probability at least $1 - e^{-t}$,
\[
\|D\|_2 \leq \sqrt{d + 2\sqrt{dt} + 2t}.
\] (24)
It then can be verified by union bound that with probability at least $1 - e^{-\frac{n^2\|S\|^2}{\lambda}} - ne^{-t}$,
\[
2D(x_0 + \lambda|S|) \leq 2\sigma_{max}^2(D) \tag{25}
\]
if
\[
n \geq \left(\sqrt{d + t} + \sqrt{\frac{(d + 2\sqrt{dt} + 2t)(x_0 + \lambda|S|)}{\lambda}}\right)^2,
\]
according to (23) and (24).

\[\square\]

Lemma 2. (Properties of the subsequences with shrinking support)

(i) All the elements of each subsequence $\mathcal{X}_t$ ($t = 1, \ldots, T$) in the subsequences with shrinking support have the same support. In addition, for any $1 \leq t_1 < t_2 \leq T$ and any $x^{(k_1)} \in \mathcal{X}_{t_1}$ and $x^{(k_2)} \in \mathcal{X}_{t_2}$, we have $k_1 < k_2$, $\text{supp}(x^{(k_2)}) \subset \text{supp}(x^{(k_1)})$.

(ii) All the subsequence except for the last one, namely $\mathcal{X}_t$ ($t = 1, \ldots, T - 1$), have finite size. Moreover, $\mathcal{X}_T$ has infinite number of elements, and there exists $k_0 \geq 0$ such that $\{x^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_T$.

\[\square\]

Proof of Lemma 2. (i) For any $1 \leq t \leq T$, let $x^{(k_1)}, x^{(k_2)} \in \mathcal{X}_t$ and $k_1 \neq k_2$. If $k_1 < k_2$, then $\text{supp}(x^{(k_2)}) \subset \text{supp}(x^{(k_1)})$ according to the support shrinkage property (12). If $\text{supp}(x^{(k_2)}) \subset \text{supp}(x^{(k_1)})$, then $|\text{supp}(x^{(k_2)})| < |\text{supp}(x^{(k_1)})|$ which contradicts with the definition of $\mathcal{X}_t$ whose elements has the same support size. Similar argument holds if $k_1 > k_2$. Therefore, all the elements of each subsequence $\mathcal{X}_t$ ($t = 1, \ldots, T$) have the same support.

For any $1 \leq t_1 < t_2 \leq T$ and any $x^{(k_1)} \in \mathcal{X}_{t_1}$ and $x^{(k_2)} \in \mathcal{X}_{t_2}$, note that $k_1 \neq k_2$ and $\text{supp}(x^{(k_2)}) \neq \text{supp}(x^{(k_1)})$ since $\mathcal{X}_{t_1}$ and $\mathcal{X}_{t_2}$ have different support size. Suppose $k_1 > k_2$. According to the support shrinkage property (12), we must have $\text{supp}(x^{(k_1)}) \subset \text{supp}(x^{(k_2)})$ and it follows that $|\text{supp}(x^{(k_1)})| < |\text{supp}(x^{(k_2)})|$, which contradicts with the definition of subsequences with shrinking support. Therefore, we must have $k_1 < k_2$, and it follows that $\text{supp}(x^{(k_2)}) \subset \text{supp}(x^{(k_1)})$.

(ii) Suppose $\mathcal{X}_t$ is an infinite sequence for some $1 \leq t \leq T - 1$. We can then obtain an infinite sequence from $\mathcal{X}_t$ in the way described as follows. We first have some $x^{(k_0)} \in \mathcal{X}_t$ for some $k_0 \geq 0$ as $\mathcal{X}_t$ is nonempty.

Suppose we obtain $\{x^{(k_j)}\}_{j=0}^{\infty}$ in the first $j \geq 0$ steps with increasing indices $\{k_j\}$, i.e. $k_j' < k_j''$ if $j' < j''$. Since $\mathcal{X}_t$ is an infinite sequence, $\mathcal{X}_t \setminus \{x^{(k_j')}\}_{j=0}^{\infty}$ is still an infinite sequence. At the $(j + 1)$-th step, we can find $x^{(k_{j+1})} \in \mathcal{X}_t \setminus \{x^{(k_j')}\}_{j=0}^{\infty}$ with $k_{j+1} > k_j$. Therefore, we obtain an infinite sequence $\{x^{(k_j)}\}_{j=0}^{\infty} \subseteq \mathcal{X}_t$ with increasing indices $\{k_j\}$. The fact that $\{k_j\}$ is increasing, i.e. $k_j' < k_j''$ if $j' < j''$, indicates that $\lim_{j \to \infty} k_j = \infty$. Now we consider an arbitrary element $x^{(k)} \in \mathcal{X}_{t+1}$. Because there must exists some $j \geq 0$ such that $k \leq k_j$, according to the support shrinkage property (12), we must have $\text{supp}(x^{(k_j)}) \subseteq \text{supp}(x^{(k)})$ which indicates that $|\text{supp}(x^{(k_j)})| \leq |\text{supp}(x^{(k)})|$. On the other hand, as $x^{(k_j)} \in \mathcal{X}_t$, the definition of the subsequences with shrinking support indicates that $|\text{supp}(x^{(k_j)})| < |\text{supp}(x^{(k)})|$. This contradiction shows that each $\mathcal{X}_t$ must have finite size for $t = 1, \ldots, T - 1$. As $\{x^{(k_j)}\}_{j=0}^{\infty}$ is an infinite sequence and $\{x^{(k_j)}\}_{j=1}^{T} \cup \{x^{(k_j)}\}_{j=0}^{\infty}$ form a disjoint union of $\{x^{(k_j)}\}$, $\mathcal{X}_T$ has infinite number of elements.

According to (i), $\mathcal{X}_T$ is an infinite sequence. By the argument in the proof of (i), there exists an infinite sequence $\{x^{(k_j)}\}_{j=0}^{\infty} \subseteq \mathcal{X}_T$, $\{k_j\}$ is increasing, and $\lim_{j \to \infty} k_j = \infty$.

For any $k > k_0$, there must exist $k_j$ with $j' \geq 1$ such that $k_{j'-1} < k \leq k_j$. According to the support shrinkage property (12),
\[
\text{supp}(x^{(k_j)}) = S^* \subseteq \text{supp}(x^{(k)}) \subseteq \text{supp}(x^{(k_j-1)}) = S^*
\]
Therefore, $|\text{supp}(x^{(k)})| = |S^*|$ and it follows that $x^{(k)} \in \mathcal{X}_T$ for any $k \geq k_0$, namely $\{x^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_T$.

\[\square\]

Denote by $S^*$ the support of any element in $\mathcal{X}_T$. If $\{x^{(k)}\}_{k=k_0}^{\infty}$ generated by Algorithm 1 has a limit point $x^*$, then the following theorem shows that the sequence $\{x^{(k)}\}_{k=k_0}^{\infty}$ converges to $x^*$, and $x^*$ is a critical point of $F(\cdot)$ whose support is $S^*$.

Theorem 2. (Convergence of PGD for the $\ell^0$ regularizer problem (1)) Suppose $s \leq \min\left\{\frac{2}{m}, \frac{2}{3}\right\}$, and $x^*$ is a limit point of $\{x^{(k)}\}_{k=k_0}^{\infty}$. Then the sequence $\{x^{(k)}\}_{k=k_0}^{\infty}$ generated by Algorithm 1 converges to $x^*$, and $x^*$ is a critical point of $F(\cdot)$. Moreover, there exists $k_0 \geq 0$ such that for all $m \geq k_0$,
\[
F(x^{(m+1)}) - F(x^*) \leq \frac{1}{2s(m - k_0 + 1)}\|x^{(k_0)} - x^*\|^2_2. \tag{26}
\]

Proof of Theorem 2. Because $x^*$ is a limit point of $\{x^{(k)}\}_{k=k_0}^{\infty}$, there must have a subsequence $\{x^{(k_j)}\}$ such that $x^{(k_j)} \to x^*$ as $j \to \infty$. In addition, $x^*$ is a limit point of $\{x^{(k)}\}_{k=k_0}$ and $F(x^*) = \inf_{k \geq 0} \{F(x^{(k)})\}$. We now show that $\text{supp}(x^*) = S^*$. To see this, we
first have \( \text{supp}(x^*) \subseteq \mathbf{S}^* \). Otherwise, pick arbitrary \( i \in \text{supp}(x^*) \setminus \mathbf{S}^* \), then \( \|x_i^{(k_j)} - x^*\|_2 \geq \|x_i^*\|_2 \), contradicting with fact that \( x_i^{(k_j)} \to x^* \).

Moreover, suppose \( \text{supp}(x^*) \subseteq \mathbf{S}^* \), we then pick arbitrary \( i \in \mathbf{S}^* \setminus \text{supp}(x^*) \). It can be shown that \( x_i^{(k_j)} \to 0 \). Otherwise, there exists \( \varepsilon > 0 \), for any \( j \), there exists \( j' \geq j \) such that \( |x_i^{(k_{j'})}| \geq \varepsilon \). It follows that \( \|x_i^{(k_{j'})} - x_i^*\|_2 \geq |x_i^{(k_{j'})}| \geq \varepsilon \), contradicting with the fact that \( x_i^{(k_j)} \to x_i^* \).

Let \( \varepsilon > 0 \) be a sufficiently small positive number such that \( sG + \varepsilon < \sqrt{2\lambda s} \). Since \( x_i^{(k_j)} \to 0 \), there exists sufficiently large \( j \) such that \( |x_i^{(k_j)}| < \varepsilon \). Let \( x_i^{(k_j+1)} = x_i^{(k_j)} - s\nabla g(x_i^{(k_j)}) \), then

\[
|x_i^{(k_j+1)}| \leq |x_i^{(k_j)}| + sG < \varepsilon + sG \leq \sqrt{2\lambda s}.
\]

It follows that \( x_i^{(k_j+1)} = 0 \) according to the update rule (2), so that \( \text{supp}(x^{(k_j+1)}) \subseteq \text{supp}(x^{(k_j)}) \setminus \{i\} \). On the other hand, note that \( x_i^{(k_j+1)} \in \mathcal{X}_i \), so we have \( \text{supp}(x^{(k_j+1)}) = \text{supp}(x^{(k_j)}) \) by Lemma 2. This contradiction shows that \( \text{supp}(x^*) \subseteq \mathbf{S}^* \) cannot hold. Therefore, \( \text{supp}(x^*) = \mathbf{S}^* \).

According to Lemma 2, there exists \( k_0 \geq 0 \) such that \( \{x_i^{(k_j)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_i \). We will prove that \( \{x_i^{(k_j)}\}_{k=k_0}^{\infty} \) converges to \( x_i^* \) in the sequel.

It follows that for any \( u, v \),

\[
g(v) \leq g(u) + \langle \nabla g(u), v - u \rangle + \frac{L}{2} \|v - u\|^2_2. \tag{27}
\]

Due to the convexity of \( g \), for any \( v \in \mathbb{R}^n \) and \( k \geq 0 \),

\[
g(x^{(k+1)}) + \langle \nabla g(x^{(k+1)}), v - x^{(k+1)} \rangle \leq g(v). \tag{28}
\]

In addition, we have

\[
x^{(k+1)} = \text{prox}_{s\lambda}(x^{(k)} - s\nabla g(x^{(k)})) = \arg\min_{x \in \mathbb{R}^d} \frac{1}{2s} \|v - (x^{(k)} - s\nabla g(x^{(k)}))\|^2_2 + h(v). \tag{29}
\]

It follows from (29) that

\[
\frac{1}{s} (x^{(k+1)} - (x^{(k)} - s\nabla g(x^{(k)}))) + \partial h(x^{(k+1)}) = 0,
\]

\[
\Rightarrow -\nabla g(x^{(k)}) - \frac{1}{s} (x^{(k+1)} - x^{(k)}) \in \partial h(x^{(k+1)}). \tag{30}
\]

Since \( x^{(k+1)} = T_{y \leq \infty}(x^{(k)} - s\nabla g(x^{(k)})) \), we have \( \partial h(x^{(k+1)}) = \partial h(x^{(k)}) \) for any \( j \in \text{supp}(x^{(k+1)}) \). It follows that for any vector \( v \in \mathbb{R}^d \) such that \( \text{supp}(v) = \text{supp}(x^{(k+1)}) \), the following equality holds:

\[
h(v) = h(x^{(k+1)}) + \langle -\nabla g(x^{(k)}) - \frac{1}{s} (x^{(k+1)} - x^{(k)}) , v - x^{(k+1)} \rangle. \tag{31}
\]

Based on (27) and (28), for any \( k \geq k_0 \) and arbitrary \( v \in \mathbb{R}^d \) we have

\[
F(x^{(k+1)}) = g(x^{(k+1)}) + h(x^{(k+1)}) \leq g(x^{(k)}) + \langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2 + h(x^{(k+1)})
\]

\[
= g(v) + \langle \nabla g(x^{(k)}), x^{(k+1)} - v \rangle + \langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \|x^{(k+1)} - v\|^2 + h(x^{(k+1)})
\]

\[
= g(v) + \langle \nabla g(x^{(k)}), x^{(k+1)} - v \rangle + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2 + h(x^{(k+1)}).
\]

When \( \text{supp}(v) = \text{supp}(x^{(k+1)}) \), according to (31) and (32),

\[
F(x^{(k+1)}) \leq g(v) + \langle \nabla g(x^{(k)}), x^{(k+1)} - v \rangle + \frac{L}{2} \|x^{(k+1)} - v\|^2 + h(x^{(k+1)})
\]

\[
= g(v) + \langle \nabla g(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2 + h(x^{(k+1)})
\]

\[
= F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k+1)} \rangle + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2
\]

\[
\leq F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k+1)} \rangle + \frac{1}{s} \|x^{(k+1)} - x^{(k)}\|^2 + \frac{L}{2} \|x^{(k+1)} - x^{(k)}\|^2
\]

\[
= F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k)} \rangle + \frac{1}{s} \|x^{(k+1)} - x^{(k)}\|^2
\]

\[
\leq F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k)} \rangle + \frac{1}{s} \|x^{(k+1)} - x^{(k)}\|^2
\]

\[
= F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k)} \rangle + \frac{1}{s} \|x^{(k+1)} - x^{(k)}\|^2
\]

\[
\leq F(v) + \frac{1}{s} \langle x^{(k+1)} - x^{(k)}, v - x^{(k)} \rangle + \frac{1}{s} \|x^{(k+1)} - x^{(k)}\|^2
\]

Now \( \text{supp}(x^*) = \text{supp}(x^{(k+1)}) = \mathbf{S}^* \), we can let \( v = x^* \) in (33), leading to

\[
F(x^{(k+1)}) - F(x^*) \leq \frac{1}{s} \langle x^{(k+1)} - x^*, x^* - x^{(k)} \rangle + \frac{1}{2s} \|x^{(k+1)} - x^*\|^2
\]

\[
\leq \frac{1}{2s} \|x^* - x^*\|^2 - \|x^{(k+1)} - x^*\|^2. \tag{34}
\]

Summing (34) over \( k = k_0, \ldots, m \) with \( m \geq k_0 \),

\[
\sum_{k=k_0}^{m} F(x^{(k+1)}) - F(x^*)
\]
\[
\sum_{k=k_0}^{m} \frac{1}{2s} \left( \|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2 \right) \\
= \frac{1}{2s} \left( \|x^{(k_0)} - x^*\|_2^2 - \|x^{(m+1)} - x^*\|_2^2 \right). \tag{35}
\]

Since \( \{F(x^{(k)})\}_k \) is non-increasing, we have
\[
\sum_{k=k_0}^{m} F(x^{(k+1)}) - F(x^*) > (m-k_0+1) F(x^{(m+1)}) - F(x^*). \]

It follows from (35) that
\[
F(x^{(m+1)}) - F(x^*) \\
\leq \frac{1}{2s(m-k_0+1)} \left( \|x^{(k_0)} - x^*\|_2^2 - \|x^{(m+1)} - x^*\|_2^2 \right) \\
\leq \frac{1}{2s(m-k_0+1)} \|x^{(k_0)} - x^*\|_2^2. \tag{36}
\]

Now we show that \( x^* \) is a critical point of \( F(\cdot) \). It follows from (30) that
\[
-\nabla g(x^{(k_j-1)}) + \frac{1}{s} (x^{(k_j)} - x^{(k_j-1)}) \in \partial h(x^{(k_j)}), \quad j \geq 1.
\]

In addition, since \( \partial F(x^{(k_j)}) = \nabla g(x^{(k_j)}) + \partial h(x^{(k_j)}) \), we have
\[
\nabla g(x^{(k_j)}) - \nabla g(x^{(k_j-1)}) - \frac{1}{s} (x^{(k_j)} - x^{(k_j-1)}) \in \partial F(x^{(k_j)}). \tag{37}
\]

Due to the fact that \( \|x^{(k_j)} - x^{(k_j-1)}\|_2 \to 0 \) as \( k \to \infty \), when \( j \to \infty \) we have
\[
\|\nabla g(x^{(k_j)}) - \nabla g(x^{(k_j-1)}) - \frac{1}{s} (x^{(k_j)} - x^{(k_j-1)})\|_2 \\
\leq L \|x^{(k_j)} - x^{(k_j-1)}\|_2 + \frac{1}{s} \|x^{(k_j)} - x^{(k_j-1)}\|_2 \\
\to 0. \tag{38}
\]

Also, as \( j \to \infty \),
\[
F(x^{(k_j)}) = g(x^{(k_j)}) + h(x^{(k_j)}) = g(x^{(k_j)}) + \lambda |S^*| \\
g(x^*) + \lambda |S^*| = F(x^*). \tag{39}
\]

Based on (37), (38) and (39), \( 0 \in \partial F(x^*) \) and \( x^* \) is a critical point of \( F(\cdot) \).

In addition, \( k_0 \) is upper bounded. Note that the sequence experiences only a finite number (at most \(|S|\)) of strict support shrinkages. The iterations of PGD between two consecutive strict support shrinkages are equivalent to those of regular gradient descent on \( g \). Suppose the last support shrinkage happens in \( k_1 \)-th iteration with \( k_1 \geq 0 \), and let \( S_1 = \text{supp}(x^{(k_1)}) \). Let \( x^* \) be the solution to the problem \( \min_{x \in S_1} g(x) \). Let the \( q \)-th \((q \in S_1)\) element of the variable incurs support shrinkage, and \( \{x^{(k)}\} \) be the sequence generated by performing gradient descent on \( g \) starting with \( x^{(k_1)} \). We can always choose \( s \) such that \( \sqrt{2\lambda s} \neq |x_q^*| \). Because \( \{x^{(k)}\} \)
converges to \( x^* \), the support shrinkage at the \( q \)-th element of the variable must happen within finite iterations. To see this, since \( \sqrt{2\lambda s} \neq |x_q^*| \), there exists a small \( \delta > 0 \) such that \( (x_q^* - \delta, x_q^* + \delta) \subset (\sqrt{2\lambda s}, \sqrt{2\lambda s}) \) or \( (x_q^* - \delta, x_q^* + \delta) \subset [-\sqrt{2\lambda s}, \sqrt{2\lambda s}] \), where \( A^c \) is the complement set of \( A \). Since \( \{x^{(k)}\} \) converges to \( x^* \), after \( T \) iterations \( \{x^{(k)}\}_{k=T} \) must fall in \( (x_q^* - \delta, x_q^* + \delta) \). If \( (x_q^* - \delta, x_q^* + \delta) \subset (\sqrt{2\lambda s}, \sqrt{2\lambda s}) \), then support shrinkage happens after \( T \) iterations. If \( (x_q^* - \delta, x_q^* + \delta) \subset [-\sqrt{2\lambda s}, \sqrt{2\lambda s}] \), support shrinkage must happen within \( T \) iterations, otherwise \( |x^{(k)}| > \sqrt{2\lambda s} \) for \( t > T \) and support shrinkage never happens at the \( q \)-th element of the variable, contradicting with the given fact. Therefore, each support shrinkage happens with finite iterations. Because shrinkage can happen at most \(|S|\) times, \( k_0 \) is upper bounded by a finite number.

\[\Box\]

**Lemma C.** For any two vectors \( u, v \in \mathbb{R}^d \), \( \|u - P_R(v)\|_2 \leq \|u - v\|_2 \) where \( \text{supp}(u) \subseteq R \).

**Proof.** We have
\[
\|u - v\|_2^2 \\
= \|P_R(u - v)\|_2^2 + \|P_{\{1, \ldots, d\}^c}(u - v)\|_2^2 \\
\geq \|P_R(u - v)\|_2^2 = \|u - P_R(v)\|_2^2. \tag{40}
\]

It follows that \( \|u - P_R(v)\|_2 \leq \|u - v\|_2 \). \[\Box\]

**Lemma 3.** (Support shrinkage for nonmonotone accelerated proximal gradient descent with support projection in Algorithm 2) The sequence \( \{x^{(k)}\}_k \) generated by Algorithm 2 satisfies
\[
\text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)}), \quad k \geq 1, \tag{41}
\]

namely the support of the sequence \( \{x^{(k)}\}_{k=1}^\infty \) shrinks.

**Proof of Lemma 3.** We prove this Lemma by mathematical induction, and we will prove that
\[
\text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)}), \quad k \geq 1. \tag{42}
\]

When \( k = 1 \), using argument similar to the proof of Lemma 1 we can show that \( \text{supp}(x^{(2)}) \subseteq \text{supp}(x^{(1)}) \), i.e. the support of \( x \) shrinks after the first iteration.

Now (42) are verified for \( k = 1 \). Suppose (42) holds for all \( k \leq k' \) with \( k' \geq 1 \). We now consider the case that \( k = k' + 1 \). Note the support projection operation in the update rule (4) for \( w^{(k)} \), and \( \text{supp}(w^{(k+1)}) \subseteq \text{supp}(x^{(k+1)}) \). Let \( q^{(k+1)} = -s \nabla g(w^{(k')}) \) and \( x^{(k'+2)} = w^{(k'+1)} - s \nabla g(w^{(k')}) \).
\( s \nabla g(w^{(k'+2)}) \). Then \( x_j^{(k'+2)} = 0 \) due to the update rule (5) for any \( j \notin \text{supp}(w^{(k+1)}) \) and
\[
|\hat{x}_j^{(k'+2)}| \leq \|q^{(k'+1)}\|_\infty \leq sG \leq \sqrt{2} sG. \quad (43)
\]
Because \( s \leq \frac{2A}{\lambda} \), the zero elements of \( w^{(k'+1)} \) remain unchanged in \( x^{(k'+2)} \), and it follows that \( \text{supp}(x^{(k'+2)}) \subseteq \text{supp}(w^{(k'+1)}) \subseteq \text{supp}(x^{(k+1)}) \). Therefore, (42) holds for \( k = k' + 1 \). It follows that (42) holds for all \( k \geq 1 \).

\[\square\]

**Theorem 3.** (Convergence of Nonmonotone Accelerated Proximal Gradient Descent for the \( \ell^0 \) regularization problem (1)) Suppose \( s \leq \min \{ \frac{2A}{\lambda}, \frac{1}{\sqrt{2}} \} \), and \( x^* \) is a limit point of \( \{x^{(k)}\}_{k=0}^{\infty} \) generated by Algorithm 2. There exists \( k_0 \geq 1 \) such that
\[ F(x^{(m+1)}) - F(x^*) \leq \frac{4}{(m+1)^2} V^{(k_0)} \quad (44) \]
for all \( m \geq k_0 \), where
\[
V^{(k_0)} \triangleq \left( \frac{1}{2s} \| (t_{k_0-1}) x^{(k_0-1)} - t_{k_0-1} x^{(k_0)} + x^* \|_2^2 + t_{k_0-1}^2 F(x^{(k_0)}) - F(x^*) \right). \quad (45)
\]

**Proof of Theorem 3.** According to Lemma 3, there exists \( k_0 \geq 0 \) such that \( \{x^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}^t \). It follows that \( \text{supp}(x^*) = S^* \).

When \( \text{supp}(v) = \text{supp}(x^{(k+1)}) \) for \( k \geq k_0 \), we have
\[
\frac{1}{2s} F(x^{(k+1)}) \leq g(v) + \langle \nabla g(w^{(k)}), x^{(k+1)} - v \rangle + \frac{L}{2} \| x^{(k+1)} - w^{(k)} \|_2^2 + h(x^{(k+1)})
\]
\[
= g(v) + \langle \nabla g(w^{(k)}), x^{(k+1)} - v \rangle + \frac{L}{2} \| x^{(k+1)} - w^{(k)} \|_2^2 + h(v)
\]
\[
+ \langle \nabla g(w^{(k)}), 1_s (x^{(k+1)} - w^{(k)}), v - x^{(k+1)} \rangle
\]
\[
= F(v) + \frac{1}{s} \langle x^{(k+1)} - w^{(k)}, v - x^{(k+1)} \rangle + \frac{L}{2} \| x^{(k+1)} - w^{(k)} \|_2^2
\]
\[
\leq F(v) + \frac{1}{s} \langle x^{(k+1)} - w^{(k)}, v - w^{(k)} \rangle - \frac{1}{8} \| x^{(k+1)} - w^{(k)} \|_2^2 + \frac{L}{2} \| x^{(k+1)} - w^{(k)} \|_2^2
\]
\[
\leq F(v) + \frac{1}{s} \langle x^{(k+1)} - w^{(k)}, v - w^{(k)} \rangle - \frac{1}{8} \| x^{(k+1)} - w^{(k)} \|_2^2
\]
\[
\leq F(v) - \frac{L}{2} \| x^{(k+1)} - w^{(k)} \|_2^2. \quad (46)
\]

Now using similar arguments in the proof of Lemma 3, let \( v = x^{(k)} \) and \( v = x^* \) in (46), we have
\[ F(x^{(k+1)}) \leq F(x^{(k)}) + \frac{1}{s} \langle x^{(k+1)} - w^{(k)} \rangle, \]
\[ x^{(k)} - w^{(k)} \]
\[
\leq (\frac{1}{s} - \frac{L}{2}) \| x^{(k+1)} - w^{(k)} \|_2^2, \quad (47)
\]
and
\[ F(x^{(k+1)}) \leq F(x^*) + \frac{1}{s} || x^{(k+1)} - w^{(k)} ||_2^2, \]
\[ x^* - w^{(k)} \]
\[
\leq (\frac{1}{s} - \frac{L}{2}) || x^{(k+1)} - w^{(k)} ||_2^2. \quad (48)
\]

(47)(\(t_k - 1\))+(48), we have
\[ t_k F(x^{(k+1)}) - (t_k - 1) F(x^*) \]
\[
\leq \frac{1}{s} || x^{(k+1)} - w^{(k)} ||_2^2, \]
\[ (t_k - 1)(x^{(k)} - w^{(k)}) + x^* - w^{(k)} \]
\[
- t_k \left( \frac{1}{s} - \frac{L}{2} \right) || x^{(k+1)} - w^{(k)} ||_2^2. \quad (49)
\]

Multiplying both sides of (49) by \( t_k \), since \( t_k^2 - t_k = t_{k-1}^2 \), we have
\[ t_k^2 \left( F(x^{(k+1)}) - F(x^*) \right) - t_{k-1}^2 \left( F(x^{(k)}) - F(x^*) \right) \]
\[
\leq \frac{1}{s} \langle t_k (x^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) \rangle + x^* - w^{(k)} \]
\[
- t_{k-1} \left( \frac{1}{s} - \frac{L}{2} \right) || x^{(k+1)} - w^{(k)} ||_2^2 \]
\[
+ x^* - w^{(k)} \]
\[
- \frac{1}{2s} \langle t_k (x^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) \rangle
\]
\[
= \frac{1}{2s} \left( || x^{(k)} - t_k x^{(k+1)} + x^* ||_2^2 
\]
\[
- || x^{(k)} - t_{k-1} x^{(k+1)} + x^* ||_2^2 \right). \quad (50)
\]

Since \( w^{(k)} = P_{\text{supp}(x^{(k)})} u^{(k)} \), it follows that \( t_k - 1)x^{(k)} - t_k P_{\text{supp}(x^{(k)})} u^{(k)} + x^* = (t_k - 1)x^{(k)} - t_k w^{(k)} + x^* \). By Lemma C and (50), we have
\[ t_k^2 \left( F(x^{(k+1)}) - F(x^*) \right) - t_{k-1}^2 \left( F(x^{(k)}) - F(x^*) \right) \]
\[
\leq \frac{1}{2s} \left( || x^{(k)} - t_k x^{(k+1)} + x^* ||_2^2 
\]
\[
- || x^{(k)} - t_{k-1} x^{(k+1)} + x^* ||_2^2 \right). \quad (51)
\]

Define \( U^{(k+1)} = (t_k - 1)x^{(k)} - t_k x^{(k+1)} + x^* \), then \( U^{(k)} = (t_k - 1)x^{(k-1)} - t_{k-1} x^{(k)} + x^* \). It can be verified that \( U^{(k)} = (t_k - 1)x^{(k)} - t_k u^{(k)} + x^* \) according to the update rule (3) for \( u^{(k)} \). Then according to (51), we have
\[ t_k^2 \left( F(x^{(k+1)}) - F(x^*) \right) - t_{k-1}^2 \left( F(x^{(k)}) - F(x^*) \right) \]
\[
\leq \frac{1}{2s} \left( || U^{(k)} ||_2^2 - || U^{(k+1)} ||_2^2 \right). \quad (52)
\]

Summing (52) over \( k = k_0, k_0 + 1, \ldots, m \) for \( m \geq k_0 \), we have
\[ t_m^2 \left( F(x^{(m+1)}) - F(x^*) \right) - t_{k_0-1}^2 \left( F(x^{(k_0)}) - F(x^*) \right) \]
\[ \leq \frac{1}{2s} \left( \| U^{(k_0)} \|^2_2 - \| U^{(m+1)} \|^2_2 \right) \]
\[ \leq \frac{1}{2s} \left( \| U^{(k_0)} \|^2_2 \right) \]
\[ = \frac{1}{2s} \left\| (t_{k-1} - 1)x^{(k_0-1)} - t_{k-1}x^{(k_0)} + x^* \right\|^2_2. \]  
(53)

It follows from (53) that
\[ F(x^{(m+1)}) - F(x^*) \]
\[ \leq \frac{1}{2s \max_t} \left( \| (t_{k-1} - 1)x^{(k_0-1)} - t_{k-1}x^{(k_0)} + x^* \right\|^2_2 \]
\[ + \frac{t_{k-1} - 1}{t_{k_0} - 1} \left( F(x^{(k_0)}) - F(x^*) \right) \]
\[ < \frac{1}{2s \max_t} \left( \| (t_{k-1} - 1)x^{(k_0-1)} - t_{k-1}x^{(k_0)} + x^* \right\|^2_2 \]
\[ + \frac{t_{k-1}^2}{t_{k_0}^2} \left( F(x^{(k_0)}) - F(x^*) \right) \]
\[ \leq \frac{4}{(m+1)^2} \left( \| (t_{k-1} - 1)x^{(k_0-1)} - t_{k-1}x^{(k_0)} + x^* \right\|^2_2 \]
\[ + \frac{4}{(m+1)^2} V^{(k_0)}, \]  
(54)

where the last inequality is due to the fact that \( t_k \geq \frac{k+1}{2} \) for \( k \geq 1 \).

**Lemma 4.** (Support shrinkage for accelerated proximal gradient descent with support projection in Algorithm 3) The sequence \( \{z^{(k)}\}_{k=1}^\infty \) and \( \{x^{(k)}\}_{k=1}^\infty \) generated by Algorithm 3 satisfy
\[ \text{supp}(z^{(k+1)}) \subseteq \text{supp}(z^{(k)}), \]  
(55)
\[ \text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)}), \]  
(56)

namely the support of both sequences shrinks.

**Proof of Lemma 4.** We prove this Lemma by mathematical induction, and we will prove that for all \( k \geq 1 \),
\[ \text{supp}(z^{(k+1)}) \subseteq \text{supp}(z^{(k)}). \]  
(57)

When \( k = 1 \), we first show that \( \text{supp}(z^{(2)}) \subseteq \text{supp}(z^{(1)}) \), i.e. the support of \( z^{(k)} \) shrinks after the first iteration.

It is now verified that (57) hold for \( k = 1 \). Suppose (57) holds for all \( k \leq k' \) with \( k' \geq 1 \). We now consider the case that \( k = k' + 1 \).

Let \( q^{(k'+1)} = - s \nabla g(w^{(k'+1)}) \) and \( x^{(k'+2)} = w^{(k'+1)} - s \nabla g(w^{(k'+1)}) \). Then \( x^{(k'+2)} = 0 \) due to the update rule (9) for any \( j \notin \text{supp}(w^{(k'+1)}) \) and \[ |x^{(k'+2)}_j| \leq s G \leq \sqrt{2 \lambda s}. \]  
(58)

Because \( s \leq \frac{2L}{\lambda s} \), the zero elements of \( w^{(k'+1)} \) remain unchanged in \( z^{(k'+2)} \). According to the support projection operation in (8), \( \text{supp}(w^{(k'+1)}) \subseteq \text{supp}(z^{(k'+1)}) = S' \). It follows that \( \text{supp}(z^{(k'+2)}) \subseteq \text{supp}(w^{(k'+1)}) \subseteq \text{supp}(z^{(k'+1)}) \). Therefore, (57) holds for \( k = k' + 1 \). It follows that (57) holds for all \( k \geq 1 \).

Now we prove (56), i.e. that for all \( k \geq 1 \), \( \text{supp}(x^{(k+1)}) \subseteq \text{supp}(x^{(k)}) \).

We have already shown that for all \( k \geq 1 \), \( \text{supp}(x^{(k)}) = \text{supp}(z^{(k)}) \) for some \( k \leq k' \). Note that \( x^{(k+1)} = z^{(k)} \) or \( x^{(k+1)} = x^{(k)} \). In the latter case, we trivially have \( \text{supp}(x^{(k+1)}) = \text{supp}(x^{(k)}) \). In the former case, \( \text{supp}(x^{(k+1)}) \subseteq \text{supp}(z^{(k)}) \subseteq \text{supp}(x^{(k)}) \) because \( k \leq k < k + 1 \). Therefore, (56) holds for all \( k \geq 1 \).

**Theorem 4.** (Convergence of Monotone Accelerated Proximal Gradient Descent for the \( \ell^1 \) regularization problem (1)) Suppose \( s \leq \min \left( \frac{2L}{\lambda s}, \frac{1}{2} \right) \), and \( x^* \) is a limit point of \( \{x^{(k)}\}_{k=0}^\infty \) generated by Algorithm 3. There exists \( k_0 \geq 1 \) such that
\[ F(x^{(m+1)}) - F(x^*) \leq \frac{4}{(m+1)^2} W^{(k_0)} \]  
(59)
for all \( m \geq k_0 \), where
\[ W^{(k_0)} \triangleq \frac{1}{2s \max_t} \left( \| (t_{k-1} - 1)x^{(k_0-1)} - t_{k-1}x^{(k_0)} + x^* \right\|^2_2 \]
\[ + \frac{4}{(m+1)^2} V^{(k_0)}. \]  
(60)

**Proof of Theorem 4.** According to Lemma 4, it can be verified that \( \{x^{(k)}\}^\infty_{k=0} \) forms at most \( T_1 \leq |S| + 1 \) subsequences with shrinking support \( \{X^{(k)}_{T_{1,1}}\}_{k=1}^{T_{1,1}} \), \( \{z^{(k)}\}^\infty_{k=0} \) also forms at most \( T_2 \leq |S| + 1 \) subsequences with shrinking support, denoted by \( \{Z^{(k)}_{T_{2,1}}\}_{k=1}^{T_{2,1}} \).

Based on Lemma 2, there exists \( k_1 \geq 0 \) such that \( \{x^{(k)}\}^\infty_{k=k_1} \subseteq X^{(k)}_{T_{1,1}} \). Similarly, there exists \( k_2 \geq 0 \) such that \( \{z^{(k)}\}^\infty_{k=k_2} \subseteq Z^{(k)}_{T_{2,1}} \). According to Lemma 2, Let all the elements of \( X^{(k)}_{T_{1,1}} \) have support \( S_1 \), and all the elements of \( Z^{(k)}_{T_{2,1}} \) have support \( S_2 \). We will show that \( S_1 = S_2 \). To see this, let \( k_0 = \max \{k_1, k_2\} \), then there exists \( k_0 \geq k_0 \) such that \( x^{(k_0)} = z^{(k_0)} \). Due to the fact that \( \{x^{(k)}\}^\infty_{k=k_1} \subseteq X^{(k)}_{T_{1,1}} \) and \( \{z^{(k)}\}^\infty_{k=k_2} \subseteq Z^{(k)}_{T_{2,1}} \), \( S_1 = \text{supp}(x^{(k_0)}) = \text{supp}(z^{(k_0)}) = S_2 \).

Let \( S_1 = S_2 = S^* \), then all the elements of \( \{x^{(k)}\}^\infty_{k=k_0} \) and \( \{z^{(k)}\}^\infty_{k=k_0} \) have the same support \( S^* \). It follows that \( \text{supp}(x^*) = S^* \).

When \( \text{supp}(v) = \text{supp}(z^{(k+1)}) \) with \( k \geq k_0 \), we have
\[ F(z^{(k+1)}) \leq g(v) + \langle \nabla g(w^{(k)}), z^{(k+1)} - v \rangle + \frac{Lg}{2} \| z^{(k+1)} - w^{(k+1)} \|^2_2 + h(z^{(k+1)}) \]
\[ g(v) + \nabla g(w^{(k)}) \cdot z^{(k+1)} - v \]
\[ + \frac{L g'}{2} ||z^{(k+1)} - w^{(k)}||^2 + h(v) \]
\[ + \langle \nabla g(w^{(k)}) + \frac{1}{s}(z^{(k+1)} - w^{(k)}), v - z^{(k+1)} \rangle \]
\[ = F(v) + \frac{1}{s}(z^{(k+1)} - w^{(k)}, v - z^{(k+1)}) \]
\[ + \frac{L g'}{2} ||z^{(k+1)} - w^{(k)}||^2 \]
\[ \leq F(v) + \frac{1}{s}(z^{(k+1)} - w^{(k)}, v - w^{(k)}) \]
\[ - \frac{1}{s} ||z^{(k+1)} - w^{(k)}||^2 + \frac{L g'}{2} ||z^{(k+1)} - w^{(k)}||^2 \]
\[ = F(v) + \frac{1}{s}(z^{(k+1)} - w^{(k)}, v - w^{(k)}) \]
\[ - \left( \frac{1}{s} - \frac{L g'}{2} \right) ||z^{(k+1)} - w^{(k)}||^2. \]
\[ (62) \times (t_k - 1) + (63), \text{ we have} \]
\[ t_k F(z^{(k+1)}) - (t_k - 1)F(x^{(k)}) - F(x^*) \]
\[ \leq \frac{1}{s}(z^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) + x^* - w^{(k)}) \]
\[ - t_k \left( \frac{1}{s} - \frac{L g'}{2} \right) ||z^{(k+1)} - w^{(k)}||^2. \]
\[ (64) \]

It follows that
\[ t_k F(z^{(k+1)}) - (t_k - 1)F(x^{(k)}) - F(x^*) \]
\[ \leq \frac{1}{s}(z^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) + x^* - w^{(k)}) \]
\[ - t_k \left( \frac{1}{s} - \frac{L g'}{2} \right) ||z^{(k+1)} - w^{(k)}||^2. \]
\[ (65) \]

Multiplying both sides of (65) by \( t_k \), since \( t_k^2 - t_k = t_{k-1}^2 \), we have
\[ t_k^2 (F(z^{(k+1)}) - F(x^*)) - t_{k-1}^2 (F(x^{(k)}) - F(x^*)) \]
\[ \leq \frac{1}{s}(t_k(z^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) + x^* - w^{(k)}) \]
\[ - \left( \frac{1}{s} - \frac{L g'}{2} \right) ||t_k(z^{(k+1)} - w^{(k)})||^2 \]
\[ \leq \frac{1}{s}(t_k(z^{(k+1)} - w^{(k)}), (t_k - 1)(x^{(k)} - w^{(k)}) + x^* - w^{(k)}) \]
\[ - \frac{1}{2s} ||t_k(z^{(k+1)} - w^{(k)})||^2 \]
\[ = \frac{1}{2s} (||t_k - 1||x^{(k)} - t_k w^{(k)} + x^*||^2 \]
\[ - ||(t_k - 1)x^{(k)} - t_k z^{(k+1)} + x^*||^2. \]
\[ (66) \]

Note that \( \text{supp}((t_k - 1)x^{(k)} + x^*) \subseteq S^* \) and \( (w^{(k)}) = F_{S^*}(u^{(k)}) \), according to Lemma C and (66), we have
\[ t_k^2 (F(z^{(k+1)}) - F(x^*)) - t_{k-1}^2 (F(x^{(k)}) - F(x^*)) \]
\[ \leq \frac{1}{2s} (||t_k - 1||x^{(k)} - t_k u^{(k)} + x^*||^2 \]
\[ - ||(t_k - 1)x^{(k)} - t_k z^{(k+1)} + x^*||^2. \]
\[ (67) \]

Define \( A^{(k+1)} = (t_k - 1)x^{(k)} - t_k z^{(k+1)} + x^* \), then \( A^{(k)} = (t_{k-1} - 1)x^{(k-1)} - t_{k-1} z^{(k)} + x^* \). It can be verified that \( A^{(k)} = (t_k - 1)x^{(k)} - t_k u^{(k)} + x^* \). Therefore,
\[ t_k^2 (F(z^{(k+1)}) - F(x^*)) - t_{k-1}^2 (F(x^{(k)}) - F(x^*)) \]
\[ \leq \frac{1}{2s} (||A^{(k)}||^2 - ||A^{(k+1)}||^2). \]
\[ (68) \]

Summing (68) over \( k = k_0, \ldots, m \) for \( m \geq k_0 \), we have
\[ t_n^2 (F(z^{(m+1)}) - F(x^*)) - t_{k_0-1}^2 (F(x^{(k_0)}) - F(x^*)) \]
\[ \leq \frac{1}{2s} (||A^{(k_0)}||^2 - ||A^{(m+1)}||^2) \leq \frac{1}{2s} ||A^{(k_0)}||^2 \]
\[ = \frac{1}{2s} (||(t_{k_0-1} - 1)x^{(k_0-1)} - t_{k_0-1} z^{(k_0)} + x^*||^2. \]
\[ (69) \]

Since \( t_k \geq \frac{k+1}{2} \) for \( k \geq 1 \), it follows from (69) that
\[ F(z^{(m+1)}) - F(x^*) \]
\[ \leq \frac{4}{(m+1)^2} \left( \frac{1}{2s} (||(t_{k_0-1} - 1)x^{(k_0-1)} - t_{k_0-1} z^{(k_0)} + x^*||^2 \]
\[ + t_{k_0-1}^2 (F(x^{(k_0)}) - F(x^*)) \right) \]
\[ \triangleq \frac{4}{(m+1)^2} W_{k_0}, \]
\[ (70) \]

References
