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

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

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

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathrm{prox}_{sh}(\mathbf{x}^{(k)} - s \nabla g(\mathbf{x}^{(k)})) \\ &= \arg\min_{\mathbf{v} \in \mathbb{R}^n} \frac{1}{2s} \|\mathbf{v} - (\mathbf{x}^{(k)} - s \nabla g(\mathbf{x}^{(k)}))\|_2^2 + \lambda \|\mathbf{v}\|_0 \\ &= T_{\sqrt{2\lambda s}}(\mathbf{x}^{(k)} - s \nabla g(\mathbf{x}^{(k)})), \end{aligned} \tag{2}$$

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

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$.
1: **for** $k = 0, \ldots,$ **do**
2: Update $\mathbf{x}^{(k+1)}$ according to (2)
3: **end for**
Output: Obtain the sparse solution $\hat{\mathbf{x}}$ upon the termination of the iterations.

1.2 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k)} + \frac{t_{k-1} - 1}{t_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \tag{3}$$

$$\mathbf{w}^{(k)} = \mathbf{P}_{\mathrm{supp}(\mathbf{x}^{(k)})}(\mathbf{u}^{(k)}), \tag{4}$$

$$\mathbf{x}^{(k+1)} = \mathrm{prox}_{sh}(\mathbf{w}^{(k)} - s \nabla g(\mathbf{w}^{(k)})), \tag{5}$$

$$t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}, \tag{6}$$

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

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$, $\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$.
1: **for** $k = 1, \ldots,$ **do**
2: Update $\mathbf{u}^{(k)}$, $\mathbf{w}^{(k)}$, $\mathbf{x}^{(k+1)}$, t_{k+1} according to (3), (4), (5), (6) respectively.
3: **end for**
Output: Obtain the sparse solution $\hat{\mathbf{x}}$ upon the termination of the iterations.

1.3 Monotone Accelerated Proximal Gradient Descent with Support Projection

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

$$\mathbf{w}^{(k)} = \mathbf{P}_{\mathrm{supp}(\mathbf{z}^{(k)})}(\mathbf{u}^{(k)}), \tag{8}$$

$$\mathbf{z}^{(k+1)} = \mathrm{prox}_{sh}(\mathbf{w}^{(k)} - s \nabla g(\mathbf{w}^{(k)})), \tag{9}$$

$$t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}, \tag{10}$$

$$\mathbf{x}^{(k+1)} = \begin{cases} \mathbf{z}^{(k+1)} & \text{if } F(\mathbf{z}^{(k+1)}) \le F(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k)} & \text{otherwise.} \end{cases} \tag{11}$$

2 PROOFS

Lemma 1. (Support shrinkage for proximal gradient descent in Algorithm 1 and sufficient decrease of the objective function) *If $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, then*

$$\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), k \ge 0, \tag{12}$$

namely the support of the sequence $\{\mathbf{x}^{(k)}\}_k$ shrinks. Moreover, the sequence of the objective $\{F(\mathbf{x}^{(k)})\}_k$ is

Algorithm 3 Monotone Accelerated Proximal Gradient Descent with Support Projection for the ℓ^0 Regularization Problem (1)

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$, $\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$.
1: **for** $k = 1, \ldots,$ **do**
2: Update $\mathbf{u}^{(k)}$, $\mathbf{w}^{(k)}$, $\mathbf{z}^{(k+1)}$, t_{k+1}, $\mathbf{x}^{(k+1)}$ according to (7), (8), (9), (10), and (11) respectively.
3: **end for**

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

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

$$F(\mathbf{x}^{(k+1)}) \le F(\mathbf{x}^{(k)}) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2. \tag{13}$$

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

With $k \ge 0$, we first show that $\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)})$, i.e. the support of the sequence shrinks. To see this, let $\tilde{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})$.

Since $\|\mathbf{y} - \mathbf{D}\mathbf{x}^{(k)}\|_2^2 = x_0$, let $\mathbf{q}^{(k)} = -s\nabla g(\mathbf{x}^{(k)}) = -2s(\mathbf{D}^\top\mathbf{D}\mathbf{x}^{(k)} - \mathbf{D}^\top\mathbf{y})$, then

$$|\tilde{\mathbf{x}}_j^{(k+1)}| \le \|\mathbf{q}^{(k)}\|_\infty \le sG,$$

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

Since

$$\mathbf{x}^{(k+1)} = \operatorname*{arg\,min}_{\mathbf{v} \in \mathbb{R}^d} \frac{1}{2s}\|\mathbf{v} - \tilde{\mathbf{x}}^{(k+1)}\|_2^2 + h(\mathbf{v}),$$

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

$$\begin{aligned} &\frac{1}{2s}\|\mathbf{x}^{(k+1)} - \tilde{\mathbf{x}}^{(k+1)}\|_2^2 + h(\mathbf{x}^{(k+1)}) \\ &\le \frac{1}{2s}\|s\nabla g(\mathbf{x}^{(k)})\|_2^2 + h(\mathbf{x}^{(k)}), \end{aligned} \tag{14}$$

which is equivalent to

$$\begin{aligned} &\langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rangle + \frac{1}{2s}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + h(\mathbf{x}^{(k+1)}) \\ &\le h(\mathbf{x}^{(k)}). \end{aligned} \tag{15}$$

In addition, since L is the Lipschitz constant for ∇g,

$$\begin{aligned} g(\mathbf{x}^{(k+1)}) &\le g(\mathbf{x}^{(k)}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rangle \\ &+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2. \end{aligned} \tag{16}$$

Combining (15) and (16), we have

$$\begin{aligned} &g(\mathbf{x}^{(k+1)}) + h(\mathbf{x}^{(k+1)}) \le g(\mathbf{x}^{(k)}) + h(\mathbf{x}^{(k)}) \\ &- (\frac{1}{2s} - \frac{L}{2})\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2. \end{aligned} \tag{17}$$

Now (12) and (13) hold for $k \ge 0$. Since the sequence $\{F(\mathbf{x}^{(k)})\}_k$ is deceasing with lower bound 0, it must converge. □

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 $\mathbf{a} = [a_1, a_2, \ldots a_D]^\top$, then for any $t > 0$,*

$$\Pr[Z \ge 2\|\mathbf{a}\|_2\sqrt{t} + 2\|\mathbf{a}\|_\infty t] \le e^{-t}. \tag{18}$$

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

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

Also, for any $t > 0$,

$$\begin{aligned} &\Pr[\sigma_n(\mathbf{A}) \le 1 - \sqrt{\frac{n}{m}} - t] < e^{-\frac{mt^2}{2}}, \\ &\Pr[\sigma_1(\mathbf{A}) \ge 1 + \sqrt{\frac{n}{m}} + t] < e^{-\frac{mt^2}{2}}. \end{aligned} \tag{20}$$

Theorem 1. *Suppose $\mathbf{D} \in \mathbb{R}^{d \times n}$ ($n \ge d$) 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 $1 - e^{-\frac{nt^2}{2}} - ne^{-t}$,*

$$\frac{2\lambda}{G^2} \ge \frac{1}{L} \tag{21}$$

if

$$n \ge \big(\sqrt{d} + t + \sqrt{\frac{(d + 2\sqrt{dt} + 2t)(x_0 + \lambda|\mathbf{S}|)}{\lambda}}\big)^2, \tag{22}$$

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 $1 - e^{-\frac{nt^2}{2}}$,

$$\sigma_{\max}(\mathbf{D}) > \sqrt{n} - \sqrt{d} - t. \tag{23}$$

Also, by Lemma A, for any $1 \le i \le n$ and $t > 0$, with probability at least $1 - e^{-t}$,

$$\|\mathbf{D}^i\|_2 \le \sqrt{d + 2\sqrt{dt} + 2t}. \tag{24}$$

It then can be verified by union bound that with probability at least $1 - e^{-\frac{nt^2}{2}} - ne^{-t}$,

$$\frac{2D^2(x_0 + \lambda|\mathbf{S}|)}{\lambda} \le 2\sigma_{\max}^2(\mathbf{D}) \tag{25}$$

if

$$n \ge \big(\sqrt{d} + t + \sqrt{\frac{(d + 2\sqrt{dt} + 2t)(x_0 + \lambda|\mathbf{S}|)}{\lambda}}\big)^2,$$

according to (23) and (24). □

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 $1 \le t_1 < t_2 \le T$ and any $\mathbf{x}^{(k_1)} \in \mathcal{X}_{t_1}$ and $\mathbf{x}^{(k_2)} \in \mathcal{X}_{t_2}$, we have $k_1 < k_2$, $\mathrm{supp}(\mathbf{x}^{(k_2)}) \subset \mathrm{supp}(\mathbf{x}^{(k_1)})$.

(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 $k_0 \ge 0$ such that $\{\mathbf{x}^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_T$.

Proof of Lemma 2. (i) For any $1 \le t \le T$, let $\mathbf{x}^{(k_1)}, \mathbf{x}^{(k_2)} \in \mathcal{X}_t$ and $k_1 \ne k_2$. If $k_1 < k_2$, then $\mathrm{supp}(\mathbf{x}^{(k_2)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k_1)})$ according to the support shrinkage property (12). If $\mathrm{supp}(\mathbf{x}^{(k_2)}) \subset \mathrm{supp}(\mathbf{x}^{(k_1)})$, then $|\mathrm{supp}(\mathbf{x}^{(k_2)})| < |\mathrm{supp}(\mathbf{x}^{(k_1)})|$ 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 $1 \le t_1 < t_2 \le T$ and any $\mathbf{x}^{(k_1)} \in \mathcal{X}_{t_1}$ and $\mathbf{x}^{(k_2)} \in \mathcal{X}_{t_2}$, note that $k_1 \ne k_2$ and $\mathrm{supp}(\mathbf{x}^{(k_2)}) \ne \mathrm{supp}(\mathbf{x}^{(k_1)})$ 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 $\mathrm{supp}(\mathbf{x}^{(k_1)}) \subset \mathrm{supp}(\mathbf{x}^{(k_2)})$ and it follows that $|\mathrm{supp}(\mathbf{x}^{(k_1)})| < |\mathrm{supp}(\mathbf{x}^{(k_2)})|$, which contradicts with the definition of subsequences with shrinking support. Therefore, we must have $k_1 < k_2$, and it follows that $\mathrm{supp}(\mathbf{x}^{(k_2)}) \subset \mathrm{supp}(\mathbf{x}^{(k_1)})$.

(ii) Suppose $\mathcal{X}_t$ is an infinite sequence for some $1 \le t \le T-1$. We can then obtain an infinite sequence from $\mathcal{X}_t$ in the way described as follows. We first have some $\mathbf{x}^{(k_0)} \in \mathcal{X}_t$ for some $k_0 \ge 0$ as $\mathcal{X}_t$ is nonempty. Suppose we obtain $\{\mathbf{x}^{(k'_j)}\}_{j'=0}^{j}$ in the first $j \ge 0$ steps with increasing indices $\{k'_j\}$, i.e. $k'_j < k''_j$ if $j' < j''$. Since $\mathcal{X}_t$ is an infinite sequence, $\mathcal{X}_t \setminus \{\mathbf{x}^{(k'_j)}\}_{j'=0}^{j}$ is still an infinite sequence. At the $(j+1)$-th step, we can find $\mathbf{x}^{(k_{j+1})} \in \mathcal{X}_t \setminus \{\mathbf{x}^{(k'_j)}\}_{j'=0}^{j}$ with $k_{j+1} > k_j$. Therefore, we obtain an infinite sequence $\{\mathbf{x}^{(k_j)}\}_{j=0}^{\infty} \subseteq \mathcal{X}_t$ with increasing increasing indices $\{k_j\}$. The fact that $\{k_j\}$ is increasing, i.e. $k'_j < k''_j$ if $j' < j''$, indicates that $\lim\limits_{j \to \infty} k_j = \infty$. Now we consider an arbitrary element $\mathbf{x}^{(\tilde{k})} \in \mathcal{X}_{t+1}$. Because there must exists some $j \ge 0$ such that $\tilde{k} \le k_j$, according to the support shrinkage property (12), we must have $\mathrm{supp}(\mathbf{x}^{(k_j)}) \subseteq \mathrm{supp}(\mathbf{x}^{(\tilde{k})})$ which indicates that $|\mathrm{supp}(\mathbf{x}^{(k_j)})| \le |\mathrm{supp}(\mathbf{x}^{(\tilde{k})})|$. On the other hand, as $\mathbf{x}^{(k_j)} \in \mathcal{X}_t$, the definition of the subsequences with shrinking support indicates that $|\mathrm{supp}(\mathbf{x}^{(\tilde{k})})| < |\mathrm{supp}(\mathbf{x}^{(k_j)})|$. This contradiction shows that each $\mathcal{X}_t$ must have finite size for $t = 1, \ldots, T-1$. As $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ is an infinite sequence and $\{\mathcal{X}_t\}_{t=1}^{T}$ form a disjoint cover of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$, $\mathcal{X}_T$ 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 $\{\mathbf{x}^{(k_j)}\}_{j=0}^{\infty} \subseteq \mathcal{X}_T$, $\{k_j\}$ is increasing, and $\lim\limits_{j \to \infty} k_j = \infty$. For any $k > k_0$, there must exist k'_j with $j' \ge 1$ such that $k_{j'-1} \le k \le k_{j'}$. According to the support shrinkage property (12),

$$\mathrm{supp}(\mathbf{x}^{(k_{j'})}) = \mathbf{S}^* \subseteq \mathrm{supp}(\mathbf{x}^{(k)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k_{j'-1})}) = \mathbf{S}^*$$

Therefore, $|\mathrm{supp}(\mathbf{x}^{(k)})| = |\mathbf{S}^*|$ and it follows that $\mathbf{x}^{(k)} \in \mathcal{X}_T$ for any $k \ge k_0$, namely $\{\mathbf{x}^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_T$.

□

Denote by $\mathbf{S}^*$ the support of any element in $\mathcal{X}_T$. If $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1 has a limit point $\mathbf{x}^*$, then the following theorem shows that the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to $\mathbf{x}^*$, and $\mathbf{x}^*$ is a critical point of $F(\cdot)$ whose support is $\mathbf{S}^*$.

Theorem 2. (Convergence of PGD for the ℓ^0 regularization problem (1)) *Suppose $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, and $\mathbf{x}^*$ is a limit point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$. Then the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1 converges to $\mathbf{x}^*$, and $\mathbf{x}^*$ is a critical point of $F(\cdot)$. Moreover, there exists $k_0 \ge 0$ such that for all $m \ge k_0$,*

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \le \frac{1}{2s(m - k_0 + 1)}\|\mathbf{x}^{(k_0)} - \mathbf{x}^*\|_2^2. \tag{26}$$

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

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

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

$$
\begin{aligned}
|\tilde{\mathbf{x}}_i^{(k_j+1)}| &\leq |\mathbf{x}_i^{(k_j)}| + sG \\
&< \varepsilon + sG \leq \sqrt{2\lambda s}.
\end{aligned}
$$

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

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

It follows that for any $\mathbf{u}, \mathbf{v}$,

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

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

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

In addition, we have

$$
\begin{aligned}
\mathbf{x}^{(k+1)} &= \mathrm{prox}_{sh}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})) \\
&= \arg\min_{\mathbf{v} \in \mathbb{R}^d} \frac{1}{2s}\|\mathbf{v} - (\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))\|_2^2 + h(\mathbf{v}).
\end{aligned} \tag{29}
$$

It follows from (29) that

$$
\begin{aligned}
&\frac{1}{s}(\mathbf{x}^{(k+1)} - (\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))) + \partial h(\mathbf{x}^{(k+1)}) = 0 \\
&\Rightarrow -\nabla g(\mathbf{x}^{(k)}) - \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \in \partial h(\mathbf{x}^{(k+1)}).
\end{aligned} \tag{30}
$$

Since $\mathbf{x}^{(k+1)} = T_{\sqrt{2\lambda s}}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))$, we have $[\partial h(\mathbf{x}^{(k+1)})]_j = 0$ for any $j \in \mathrm{supp}(\mathbf{x}^{(k+1)})$. It follows that for any vector $\mathbf{v} \in \mathbb{R}^d$ such that $\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\mathbf{x}^{(k+1)})$, the following equality holds:

$$
\begin{aligned}
h(\mathbf{v}) = h(\mathbf{x}^{(k+1)}) + \langle -\nabla g(\mathbf{x}^{(k)}) - \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), \\
\mathbf{v} - \mathbf{x}^{(k+1)} \rangle.
\end{aligned} \tag{31}
$$

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

$$
\begin{aligned}
&F(\mathbf{x}^{(k+1)}) = g(\mathbf{x}^{(k+1)}) + h(\mathbf{x}^{(k+1)}) \\
&\leq g(\mathbf{x}^{(k)}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + h(\mathbf{x}^{(k+1)}) \\
&\leq g(\mathbf{v}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k)} - \mathbf{v} \rangle + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \rangle \\
&+ \frac{L}{2}\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h(\mathbf{z}^{(k+1)}) \\
&= g(\mathbf{v}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \\
&+ h(\mathbf{x}^{(k+1)}).
\end{aligned} \tag{32}
$$

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

$$
\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\leq g(\mathbf{v}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v} \rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + h(\mathbf{x}^{(k+1)}) \\
&= g(\mathbf{v}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v} \rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + h(\mathbf{v}) \\
&+ \langle \nabla g(\mathbf{x}^{(k)}) + \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), \mathbf{v} - \mathbf{x}^{(k+1)} \rangle \\
&= F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{v} - \mathbf{x}^{(k+1)} \rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \\
&= F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{v} - \mathbf{x}^{(k)} \rangle \\
&- \frac{1}{s}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \\
&= F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{v} - \mathbf{x}^{(k)} \rangle \\
&- \left(\frac{1}{s} - \frac{L}{2}\right)\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \\
&\leq F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{v} - \mathbf{x}^{(k)} \rangle \\
&- \frac{1}{2s}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2.
\end{aligned} \tag{33}
$$

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

$$
\begin{aligned}
&F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \\
&\leq \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{x}^* - \mathbf{x}^{(k)} \rangle - \frac{1}{2s}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \\
&= \frac{1}{2s}\left(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|_2^2\right).
\end{aligned} \tag{34}
$$

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

$$
\sum_{k=k_0}^{m} F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*)
$$

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

Since $\{F(\mathbf{x}^{(k)})\}_k$ is non-increasing, we have $\sum\limits_{k=k_0}^{m} F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) > (m-k_0+1)F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*)$. It follows from (35) that

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*)$$
$$\leq \frac{1}{2s(m-k_0+1)}\left(\|\mathbf{x}^{(k_0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(m+1)} - \mathbf{x}^*\|_2^2\right)$$
$$\leq \frac{1}{2s(m-k_0+1)}\|\mathbf{x}^{(k_0)} - \mathbf{x}^*\|_2^2. \tag{36}$$

Now we show that $\mathbf{x}^*$ is a critical point of $F(\cdot)$. It follows from (30) that $-\nabla g(\mathbf{x}^{(k_j-1)}) - \frac{1}{s}(\mathbf{x}^{(k_j)} - \mathbf{x}^{(k_j-1)}) \in \partial h(\mathbf{x}^{(k_j)})$ for $k_j \geq 1$. In addition, since $\partial F(\mathbf{x}^{(k_j)}) = \nabla g(\mathbf{x}^{(k_j)}) + \partial h(\mathbf{x}^{(k_j)})$, we have

$$\nabla g(\mathbf{x}^{(k_j)}) - \nabla g(\mathbf{x}^{(k_j-1)}) - \frac{1}{s}(\mathbf{x}^{(k_j)} - \mathbf{x}^{(k_j-1)}) \in \partial F(\mathbf{x}^{(k_j)}). \tag{37}$$

Due to the fact that $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_2 \to 0$ as $k \to \infty$, when $j \to \infty$ we have

$$\|\nabla g(\mathbf{x}^{(k_j)}) - \nabla g(\mathbf{x}^{(k_j-1)}) - \frac{1}{s}(\mathbf{x}^{(k_j)} - \mathbf{x}^{(k_j-1)})\|_2$$
$$\leq L\|\mathbf{x}^{(k_j)}) - \mathbf{x}^{(k_j-1)}\|_2 + \frac{1}{s}\|\mathbf{x}^{(k_j)} - \mathbf{x}^{(k_j-1)}\|_2$$
$$\to 0. \tag{38}$$

Also, as $j \to \infty$,

$$F(\mathbf{x}^{(k_j)}) = g(\mathbf{x}^{(k_j)}) + h(\mathbf{x}^{(k_j)}) = g(\mathbf{x}^{(k_j)}) + \lambda|\mathbf{S}^*|$$
$$\to g(\mathbf{x}^*) + \lambda|\mathbf{S}^*| = g(\mathbf{x}^*) + h(\mathbf{x}^*) = F(\mathbf{x}^*). \tag{39}$$

Based on (37), (38) and (39), $\mathbf{0} \in \partial F(\mathbf{x}^*)$ and $\mathbf{x}^*$ 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 $|\mathbf{S}|$) 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 $\mathbf{S}_1 = \mathrm{supp}(\mathbf{x}^{(k_1)})$. Let $\mathbf{x}'$ be the solution to the problem $\min_{\mathbf{x}, \mathrm{supp}(\mathbf{x}) = \mathbf{S}_1} g(\mathbf{x})$. Let the q-th ($q \in \mathbf{S}_1$) element of the variable incurs support shrinkage, and $\{\mathbf{x}'^{(t)}\}$ be the sequence generated by performing gradient descent on g staring with $\mathbf{x}^{(k_1)}$. We can always choose s such that $\sqrt{2\lambda s} \neq |\mathbf{x}'_q|$. Because $\{\mathbf{x}'^{(t)}\}$ converges to $\mathbf{x}'$, the support shrinkage at the q-th element of the variable must happen within finite iterations. To see this, since $\sqrt{2\lambda s} \neq \mathbf{x}'_q$, there exists a small $\delta > 0$ such that $(\mathbf{x}'_q - \delta, \mathbf{x}'_q + \delta) \subset (-\sqrt{2\lambda s}, \sqrt{2\lambda s})$ or $(\mathbf{x}'_q - \delta, \mathbf{x}'_q + \delta) \subset [-\sqrt{2\lambda s}, \sqrt{2\lambda s}]^{\complement}$, where $\mathbf{A}^{\complement}$ is the complement set of $\mathbf{A}$. Since $\{\mathbf{x}'^{(t)}\}$ converges to $\mathbf{x}'$, after T iterations $\{\mathbf{x}'^{(t)}\}_{t>T}$ must fall in $(\mathbf{x}'_q - \delta, \mathbf{x}'_q + \delta)$. If $(\mathbf{x}'_q - \delta, \mathbf{x}'_q + \delta) \subset (-\sqrt{2\lambda s}, \sqrt{2\lambda s})$, then support shrinkage happens after T iterations. If $(\mathbf{x}'_q - \delta, \mathbf{x}'_q + \delta) \subset [-\sqrt{2\lambda s}, \sqrt{2\lambda s}]^{\complement}$, support shrinkage must happen within T iterations, otherwise $|\mathbf{x}'^{(t)}| > \sqrt{2\lambda s}$ 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 $|\mathbf{S}|$ times, k_0 is upper bounded by a finite number.

□

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

Proof. We have

$$\|\mathbf{u} - \mathbf{v}\|_2^2$$
$$= \|\mathbf{P}_{\mathbf{R}}(\mathbf{u} - \mathbf{v})\|_2^2 + \|\mathbf{P}_{\{1,\ldots,d\}\setminus\mathbf{R}}(\mathbf{u} - \mathbf{v})\|_2^2$$
$$\geq \|\mathbf{P}_{\mathbf{R}}(\mathbf{u} - \mathbf{v})\|_2^2 = \|\mathbf{u} - \mathbf{P}_{\mathbf{R}}(\mathbf{v})\|_2^2. \tag{40}$$

It follows that $\|\mathbf{u} - \mathbf{P}_{\mathbf{R}}(\mathbf{v})\|_2 \leq \|\mathbf{u} - \mathbf{v}\|_2$. □

Lemma 3. (Support shrinkage for nonmonotone accelerated proximal gradient descent with support projection in Algorithm 2) *The sequence* $\{\mathbf{x}^{(k)}\}_k$ *generated by Algorithm 2 satisfies*

$$\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), k \geq 1, \tag{41}$$

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

Proof of Lemma 3. We prove this Lemma by mathematical induction, and we will prove that

$$\mathrm{supp}(\mathbf{x}^{(\bar{k}+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(\bar{k})}), \bar{k} \geq 1. \tag{42}$$

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

Now (42) are verified for $\bar{k} = 1$. Suppose (42) holds for all $\bar{k} \leq k'$ with $k' \geq 1$. We now consider the case that $\bar{k} = k' + 1$.

Note the support projection operation in the update rule (4) for $\mathbf{w}^{(k)}$, and $\mathrm{supp}(\mathbf{w}^{(k'+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k'+1)})$. Let $\mathbf{q}^{(k'+1)} = -s\nabla g(\mathbf{w}^{(k'+1)})$ and $\tilde{\mathbf{x}}_j^{(k'+2)} = \mathbf{w}^{(k'+1)} -$

$s\nabla g(\mathbf{w}^{(k'+1)})$. Then $\mathbf{x}_j^{(k'+2)} = 0$ due to the update rule (5) for any $j \notin \text{supp}(\mathbf{w}^{(k'+1)})$ and

$$|\tilde{\mathbf{x}}_j^{(k'+2)}| \le \|\mathbf{q}^{(k'+1)}\|_\infty \le sG \le \sqrt{2\lambda s}. \tag{43}$$

Because $s \le \frac{2\lambda}{G^2}$, the zero elements of $\mathbf{w}^{(k'+1)}$ remain unchanged in $\mathbf{x}^{(k'+2)}$, and it follows that $\text{supp}(\mathbf{x}^{(k'+2)}) \subseteq \text{supp}(\mathbf{w}^{(k'+1)}) \subseteq \text{supp}(\mathbf{x}^{(k'+1)})$. Therefore, (42) holds for $\bar{k} = k'+1$. It follows that (42) holds for all $\bar{k} \ge 1$.

□

Theorem 3. (Convergence of Nonmonotone Accelerated Proximal Gradient Descent for the ℓ^0 regularization problem (1)) *Suppose* $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, *and* $\mathbf{x}^*$ *is a limit point of* $\{\mathbf{x}^{(k)}\}_{k=0}^\infty$ *generated by Algorithm 2. There exists* $k_0 \ge 1$ *such that*

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \le \frac{4}{(m+1)^2} V^{(k_0)} \tag{44}$$

for all $m \ge k_0$, *where*

$$V^{(k_0)} \triangleq \Big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2 + t_{k_0-1}^2\big(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\big)\Big). \tag{45}$$

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

When $\text{supp}(\mathbf{v}) = \text{supp}(\mathbf{x}^{(k+1)})$ for $k \ge k_0$, we have

$$\begin{aligned}
&F(\mathbf{x}^{(k+1)}) \le g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v}\rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h(\mathbf{x}^{(k+1)}) \\
&= g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v}\rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h(\mathbf{v}) \\
&+ \langle \nabla g(\mathbf{w}^{(k)}) + \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}), \mathbf{v} - \mathbf{x}^{(k+1)}\rangle \\
&= F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{x}^{(k+1)}\rangle \\
&+ \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 \\
&\le F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)}\rangle \\
&- \frac{1}{s}\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 \\
&= F(\mathbf{v}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)}\rangle \\
&- \big(\frac{1}{s} - \frac{L}{2}\big)\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2.
\end{aligned} \tag{46}$$

Now using similar arguments in the proof of Lemma 3, let $\mathbf{v} = \mathbf{x}^{(k)}$ and $\mathbf{v} = \mathbf{x}^*$ in in (46), we have

$$\begin{aligned}
&F(\mathbf{x}^{(k+1)}) \le F(\mathbf{x}^{(k)}) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, \\
&\mathbf{x}^{(k)} - \mathbf{w}^{(k)}\rangle - \big(\frac{1}{s} - \frac{L}{2}\big)\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2,
\end{aligned} \tag{47}$$

and

$$\begin{aligned}
&F(\mathbf{x}^{(k+1)}) \le F(\mathbf{x}^*) + \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, \\
&\mathbf{x}^* - \mathbf{w}^{(k)}\rangle - \big(\frac{1}{s} - \frac{L}{2}\big)\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2.
\end{aligned} \tag{48}$$

(47)$\times(t_k - 1)+$ (48), we have

$$\begin{aligned}
&t_k F(\mathbf{x}^{(k+1)}) - (t_k - 1)F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*) \\
&\le \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}, (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)}\rangle \\
&- t_k\big(\frac{1}{s} - \frac{L}{2}\big)\|\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2.
\end{aligned} \tag{49}$$

Multiplying both sides of (49) by t_k, since $t_k^2 - t_k = t_{k-1}^2$, we have

$$\begin{aligned}
&t_k^2\big(F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{s}\langle t_k(\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}), (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \\
&\mathbf{x}^* - \mathbf{w}^{(k)}\rangle - \big(\frac{1}{s} - \frac{L}{2}\big)\|t_k(\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)})\|_2^2 \\
&\le \frac{1}{s}\langle t_k(\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)}), (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) \\
&+ \mathbf{x}^* - \mathbf{w}^{(k)}\rangle - \frac{1}{2s}\|t_k(\mathbf{x}^{(k+1)} - \mathbf{w}^{(k)})\|_2^2 \\
&= \frac{1}{2s}\big(\|(t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{w}^{(k)} + \mathbf{x}^*\|_2^2 \\
&- \|(t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{x}^{(k+1)} + \mathbf{x}^*\|_2^2\big).
\end{aligned} \tag{50}$$

Since $\mathbf{w}^{(k)} = \mathbf{P}_{\text{supp}(\mathbf{x}^{(k)})}(\mathbf{u}^{(k)})$, it follows that $(t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{P}_{\text{supp}(\mathbf{x}^{(k)})}(\mathbf{u}^{(k)}) + \mathbf{x}^* = (t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{w}^{(k)} + \mathbf{x}^*$. By Lemma C and (50), we have

$$\begin{aligned}
&t_k^2\big(F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{2s}\big(\|(t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{u}^{(k)} + \mathbf{x}^*\|_2^2 \\
&- \|(t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{x}^{(k+1)} + \mathbf{x}^*\|_2^2\big).
\end{aligned} \tag{51}$$

Define $\mathbf{U}^{(k+1)} = (t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{x}^{(k+1)} + \mathbf{x}^*$, then $\mathbf{U}^{(k)} = (t_{k-1} - 1)\mathbf{x}^{(k-1)} - t_{k-1}\mathbf{x}^{(k)} + \mathbf{x}^*$. It can be verified that $\mathbf{U}^{(k)} = (t_k - 1)\mathbf{x}^{(k)} - t_k\mathbf{u}^{(k)} + \mathbf{x}^*$ according to the update rule (3) for $\mathbf{u}^{(k)}$. Then according to (51), we have

$$\begin{aligned}
&t_k^2\big(F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{2s}\big(\|\mathbf{U}^{(k)}\|_2^2 - \|\mathbf{U}^{(k+1)}\|_2^2\big).
\end{aligned} \tag{52}$$

Summing (52) over $k = k_0, k_0, +1, \ldots, m$ for $m \ge k_0$, we have

$$t_m^2\big(F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*)\big) - t_{k_0-1}^2\big(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\big)$$

$$
\begin{aligned}
&\le \frac{1}{2s}\left(\|\mathbf{U}^{(k_0)}\|_2^2 - \|\mathbf{U}^{(m+1)}\|_2^2\right) \\
&\le \frac{1}{2s}\|\mathbf{U}^{(k_0)}\|_2^2 \\
&= \frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2. \qquad (53)
\end{aligned}
$$

It follows from (53) that

$$
\begin{aligned}
&F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \\
&\le \frac{1}{2st_m^2}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2 \\
&+ \frac{t_{k_0-1}^2}{t_m^2}\left(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\right) \\
&< \frac{1}{t_m^2}\Big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2 \\
&+ t_{k_0-1}^2\left(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\right)\Big) \\
&\le \frac{4}{(m+1)^2}\Big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2 \\
&+ t_{k_0-1}^2\left(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\right)\Big) \\
&\triangleq \frac{4}{(m+1)^2}V^{(k_0)}, \qquad (54)
\end{aligned}
$$

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

$\square$

Lemma 4. (Support shrinkage for accelerated proximal gradient descent with support projection in Algorithm 3) *The sequence $\{\mathbf{z}^{(k)}\}_{k=1}^{\infty}$ and $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ generated by Algorithm 3 satisfy*

$$
\mathrm{supp}(\mathbf{z}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{z}^{(k)}), \qquad (55)
$$
$$
\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), \qquad (56)
$$

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 $\bar{k} \ge 1$,

$$
\mathrm{supp}(\mathbf{z}^{(\bar{k}+1)}) \subseteq \mathrm{supp}(\mathbf{z}^{(\bar{k})}). \qquad (57)
$$

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

It is now verified that (57) hold for $\bar{k} = 1$. Suppose (57) holds for all $\bar{k} \le k'$ with $k' \ge 1$. We now consider the case that $\bar{k} = k' + 1$.

Let $\mathbf{q}^{(k'+1)} = -s\nabla g(\mathbf{w}^{(k'+1)})$ and $\tilde{\mathbf{x}}_j^{(k'+2)} = \mathbf{w}^{(k'+1)} - s\nabla g(\mathbf{w}^{(k'+1)})$. Then $\mathbf{x}_j^{(k'+2)} = 0$ due to the update rule (9) for any $j \notin \mathrm{supp}(\mathbf{w}^{(k'+1)})$ and

$$
|\tilde{\mathbf{x}}_j^{(k'+2)}| \le sG \le \sqrt{2\lambda s}. \qquad (58)
$$

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

Now we prove (56), i.e. that for all $k \ge 1$, $\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)})$.

We have already shown that for all $k \ge 1$, $\mathrm{supp}(\mathbf{x}^{(k)}) = \mathrm{supp}(\mathbf{z}^{(\tilde{k})})$ for some $\tilde{k} \le k$. Note that $\mathbf{x}^{(k+1)} = \mathbf{z}^{(k+1)}$ or $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)}$. In the latter case, we trivially have $\mathrm{supp}(\mathbf{x}^{(k+1)}) = \mathrm{supp}(\mathbf{x}^{(k)})$. In the former case, $\mathrm{supp}(\mathbf{x}^{(k+1)}) = \mathrm{supp}(\mathbf{z}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{z}^{(\tilde{k})}) = \mathrm{supp}(\mathbf{x}^{(k)})$ because $\tilde{k} \le k < k + 1$. Therefore, (56) holds for all $k \ge 1$. $\square$

Theorem 4. (Convergence of Monotone Accelerated Proximal Gradient Descent for the ℓ^0 regularization problem (1)) *Suppose $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, and $\mathbf{x}^*$ is a limit point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 3. There exists $k_0 \ge 1$ such that*

$$
F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \le \frac{4}{(m+1)^2}W^{(k_0)} \qquad (59)
$$

for all $m \ge k_0$, where

$$
\begin{aligned}
W^{(k_0)} &\triangleq \big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\|_2^2 \\
&+ t_{k_0-1}^2(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*))\big). \qquad (60)
\end{aligned}
$$

Proof of Theorem 4. According to Lemma 4, it can be verified that $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ forms at most $T_1 \le |\mathbf{S}| + 1$ subsequences with shrinking support $\{\mathcal{X}_t\}_{t=1}^{T_1}$, and $\{\mathbf{z}^{(k)}\}_{k=0}^{\infty}$ also forms at most $T_2 \le |\mathbf{S}| + 1$ subsequences with shrinking support, denoted by $\{\mathcal{Z}_t\}_{t=1}^{T_2}$.

Based on Lemma 2, there exists $k_1 \ge 0$ such that $\{\mathbf{x}^{(k)}\}_{k=k_1}^{\infty} \subseteq \mathcal{X}_{T_1}$. Similarly, there exists $k_2 \ge 0$ such that $\{\mathbf{z}^{(k)}\}_{k=k_2}^{\infty} \subseteq \mathcal{Z}_{T_2}$. According to Lemma 2, Let all the elements of $\mathcal{X}_{T_1}$ have support $\mathbf{S}_1$, and all the elements of $\mathcal{Z}_{T_2}$ have support $\mathbf{S}_2$. We will show that $\mathbf{S}_1 = \mathbf{S}_2$. To see this, let $k_0 = \max\{k_1, k_2\}$, then there exists $k' \ge k_0$ such that $\mathbf{x}^{(k')} = \mathbf{z}^{(k')}$. Due to the fact that $\{\mathbf{x}^{(k)}\}_{k=k_1}^{\infty} \subseteq \mathcal{X}_{T_1}$ and $\{\mathbf{z}^{(k)}\}_{k=k_2}^{\infty} \subseteq \mathcal{Z}_{T_2}$, $\mathbf{S}_1 = \mathrm{supp}(\mathbf{x}^{(k')}) = \mathrm{supp}(\mathbf{z}^{(k')}) = \mathbf{S}_2$.

Let $\mathbf{S}_1 = \mathbf{S}_2 = \mathbf{S}^*$, then all the elements of $\{\mathbf{x}^{(k)}\}_{k=k_0}^{\infty}$ and $\{\mathbf{z}^{(k)}\}_{k=k_0}^{\infty}$ have the same support $\mathbf{S}^*$. It follows that $\mathrm{supp}(\mathbf{x}^*) = \mathbf{S}^*$.

When $\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\mathbf{z}^{(k+1)})$ with $k \ge k_0$, we have

$$
\begin{aligned}
&F(\mathbf{z}^{(k+1)}) \le g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{z}^{(k+1)} - \mathbf{v}\rangle \\
&+ \frac{L_{\mathbf{S}'}}{2}\|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h(\mathbf{z}^{(k+1)})
\end{aligned}
$$

$$
\begin{aligned}
&= g(\mathbf{v}) + \langle \nabla g(\mathbf{w}^{(k)}), \mathbf{z}^{(k+1)} - \mathbf{v} \rangle \\
&+ \frac{L_{\mathbf{S}'}}{2} \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + h(\mathbf{v}) \\
&+ \langle \nabla g(\mathbf{w}^{(k)}) + \frac{1}{s}(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), \mathbf{v} - \mathbf{z}^{(k+1)} \rangle \\
&= F(\mathbf{v}) + \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{z}^{(k+1)} \rangle \\
&+ \frac{L_{\mathbf{S}'}}{2} \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 \\
&\le F(\mathbf{v}) + \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)} \rangle \\
&- \frac{1}{s} \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 + \frac{L_{\mathbf{S}'}}{2} \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2 \\
&= F(\mathbf{v}) + \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{v} - \mathbf{w}^{(k)} \rangle \\
&- \big(\frac{1}{s} - \frac{L_{\mathbf{S}'}}{2}\big) \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2. \qquad (61)
\end{aligned}
$$

Note that $\mathrm{supp}(\mathbf{x}^{(k)}) = \mathrm{supp}(\mathbf{x}^*) = \mathbf{S}^*$ for $k \ge k_0$. Using similar arguments in the proof of Lemma 3, let $\mathbf{v} = \mathbf{x}^{(k)}$ and $\mathbf{v} = \mathbf{x}^*$ in (61) in the proof of Lemma 4, we have

$$
\begin{aligned}
&F(\mathbf{z}^{(k+1)}) \le F(\mathbf{x}^{(k)}) + \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{x}^{(k)} - \mathbf{w}^{(k)} \rangle \\
&- \big(\frac{1}{s} - \frac{L}{2}\big) \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2, \qquad (62)
\end{aligned}
$$

and

$$
\begin{aligned}
&F(\mathbf{z}^{(k+1)}) \le F(\mathbf{x}^*) + \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, \mathbf{x}^* - \mathbf{w}^{(k)} \rangle \\
&- \big(\frac{1}{s} - \frac{L}{2}\big) \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2. \qquad (63)
\end{aligned}
$$

(62)$\times (t_k - 1)+$ (63), we have

$$
\begin{aligned}
&t_k F(\mathbf{z}^{(k+1)}) - (t_k - 1) F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*) \\
&\le \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)} \rangle \\
&- t_k \big(\frac{1}{s} - \frac{L}{2}\big) \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2. \qquad (64)
\end{aligned}
$$

It follows that

$$
\begin{aligned}
&t_k \big(F(\mathbf{z}^{(k+1)}) - F(\mathbf{x}^*)\big) - (t_k - 1)\big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{s} \langle \mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}, (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)} \rangle \\
&- t_k \big(\frac{1}{s} - \frac{L}{2}\big) \|\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}\|_2^2. \qquad (65)
\end{aligned}
$$

Multiplying both sides of (65) by t_k, since $t_k^2 - t_k = t_{k-1}^2$, we have

$$
\begin{aligned}
&t_k^2 \big(F(\mathbf{z}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2 \big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{s} \langle t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)} \rangle \\
&- \big(\frac{1}{s} - \frac{L}{2}\big) \|t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)})\|_2^2 \\
&\le \frac{1}{s} \langle t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)}), (t_k - 1)(\mathbf{x}^{(k)} - \mathbf{w}^{(k)}) + \mathbf{x}^* - \mathbf{w}^{(k)} \rangle \\
&- \frac{1}{2s} \|t_k(\mathbf{z}^{(k+1)} - \mathbf{w}^{(k)})\|_2^2 \\
&= \frac{1}{2s} \big(\|(t_k - 1)\mathbf{x}^{(k)} - t_k \mathbf{w}^{(k)} + \mathbf{x}^*\|_2^2 \\
&- \|(t_k - 1)\mathbf{x}^{(k)} - t_k \mathbf{z}^{(k+1)} + \mathbf{x}^*\|_2^2 \big). \qquad (66)
\end{aligned}
$$

Note that $\mathrm{supp}((t_k - 1)\mathbf{x}^{(k)} + \mathbf{x}^*) \subseteq \mathbf{S}^*$ and $(\mathbf{w}^{(k)}) = \mathbf{P}_{\mathbf{S}^*}(\mathbf{u}^{(k)})$, according to Lemma C and (66), we have

$$
\begin{aligned}
&t_k^2 \big(F(\mathbf{z}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2 \big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{2s} \big(\|(t_k - 1)\mathbf{x}^{(k)} - t_k \mathbf{u}^{(k)} + \mathbf{x}^*\|_2^2 \\
&- \|(t_k - 1)\mathbf{x}^{(k)} - t_k \mathbf{z}^{(k+1)} + \mathbf{x}^*\|_2^2 \big). \qquad (67)
\end{aligned}
$$

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

$$
\begin{aligned}
&t_k^2 \big(F(\mathbf{z}^{(k+1)}) - F(\mathbf{x}^*)\big) - t_{k-1}^2 \big(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{2s} \big(\|\mathbf{A}^{(k)}\|_2^2 - \|\mathbf{A}^{(k+1)}\|_2^2 \big). \qquad (68)
\end{aligned}
$$

Summing (68) over $k = k_0, \ldots, m$ for $m \ge k_0$, we have

$$
\begin{aligned}
&t_m^2 \big(F(\mathbf{z}^{(m+1)}) - F(\mathbf{x}^*)\big) - t_{k_0-1}^2 \big(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\big) \\
&\le \frac{1}{2s} \big(\|\mathbf{A}^{(k_0)}\|_2^2 - \|\mathbf{A}^{(m+1)}\|_2^2 \big) \le \frac{1}{2s} \|\mathbf{A}^{(k_0)}\|_2^2 \\
&= \frac{1}{2s} \|(t_{k_0-1} - 1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\|_2^2. \qquad (69)
\end{aligned}
$$

Since $t_k \ge \frac{k+1}{2}$ for $k \ge 1$, it follows from (69) that

$$
\begin{aligned}
&F(\mathbf{z}^{(m+1)}) - F(\mathbf{x}^*) \\
&\le \frac{4}{(m+1)^2} \big(\frac{1}{2s} \|(t_{k_0-1} - 1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\|_2^2 \\
&+ t_{k_0-1}^2 (F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)) \big) \\
&\triangleq \frac{4}{(m+1)^2} W^{(k_0)}. \qquad (70)
\end{aligned}
$$

$\square$

References

K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In Lindenstrauss, editor, *Handbook on the Geometry of Banach spaces*, volume 1, pages 317–366. Elsevier Science, 2001.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000.