A. Technical Lemmas

The following lemma is a characterization of the co-coercivity of the objective function $F(\boldsymbol{x})$. A similar result was obtained in Nguyen et al. (2014, Corollary 8) but we present a refined analysis which is essential for our purpose.

Lemma 9. *For a given support set Ω, assume that the continuous function $F(\boldsymbol{x})$ is $M_{|\Omega|}$-RSS and is m_K-RSC for some sparsity level K. Then, for all vectors $\boldsymbol{w}$ and $\boldsymbol{w}'$ with $|\mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{w}')\cup\Omega| \leq K$, we have*

$$\|\nabla_\Omega F(\boldsymbol{w}') - \nabla_\Omega F(\boldsymbol{w})\|^2 \leq 2M_{|\Omega|}\big(F(\boldsymbol{w}') - F(\boldsymbol{w}) - \langle \nabla F(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w}\rangle\big).$$

Proof. We define an auxiliary function

$$G(\boldsymbol{x}) \overset{\text{def}}{=} F(\boldsymbol{x}) - \langle \nabla F(\boldsymbol{w}), \boldsymbol{x}\rangle.$$

For all vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$\|\nabla G(\boldsymbol{x}) - \nabla G(\boldsymbol{y})\| = \|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\| \leq M_{|\mathrm{supp}(\boldsymbol{x}-\boldsymbol{y})|}\, \|\boldsymbol{x}-\boldsymbol{y}\|,$$

which is equivalent to

$$G(\boldsymbol{x}) - G(\boldsymbol{y}) - \langle \nabla G(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \leq \frac{M_r}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2, \tag{7}$$

where $r = |\mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{y})|$. On the other hand, due to the RSC property of $F(\boldsymbol{x})$, we obtain

$$G(\boldsymbol{x}) - G(\boldsymbol{w}) = F(\boldsymbol{x}) - F(\boldsymbol{w}) - \langle \nabla F(\boldsymbol{w}), \boldsymbol{x}-\boldsymbol{w}\rangle \geq \frac{m_{|\mathrm{supp}(\boldsymbol{x}-\boldsymbol{w})|}}{2}\|\boldsymbol{x}-\boldsymbol{w}\|^2 \geq 0,$$

provided that $|\mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{w})| \leq K$. For the given support set Ω, we pick $\boldsymbol{x} = \boldsymbol{w}' - \frac{1}{M_{|\Omega|}}\nabla_\Omega G(\boldsymbol{w}')$. Clearly, for such a choice of $\boldsymbol{x}$, we have $\mathrm{supp}\,(\boldsymbol{x}-\boldsymbol{w}) = \mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{w}')\cup\Omega$. Hence, by assuming that $|\mathrm{supp}\,(\boldsymbol{w}-\boldsymbol{w}')\cup\Omega|$ is not larger than K, we get

$$\begin{aligned}
G(\boldsymbol{w}) &\leq G\left(\boldsymbol{w}' - \frac{1}{M_{|\Omega|}}\nabla_\Omega G(\boldsymbol{w}')\right)\\
&\overset{(7)}{\leq} G(\boldsymbol{w}') + \left\langle \nabla G(\boldsymbol{w}'), -\frac{1}{M_{|\Omega|}}\nabla_\Omega G(\boldsymbol{w}')\right\rangle + \frac{1}{2M_{|\Omega|}}\|\nabla_\Omega G(\boldsymbol{w}')\|^2\\
&= G(\boldsymbol{w}') - \frac{1}{2M_{|\Omega|}}\|\nabla_\Omega G(\boldsymbol{w}')\|^2.
\end{aligned}$$

Now expanding $\nabla_\Omega G(\boldsymbol{w}')$ and rearranging the terms give the desired result. □

Lemma 10 (Lemma 1 in Wang et al. (2016))**.** *Let $\boldsymbol{u}$ and $\boldsymbol{z}$ be two distinct vectors and let $W = \mathrm{supp}\,(\boldsymbol{u})\cap\mathrm{supp}\,(\boldsymbol{z})$. Also, let U be the support set of the top r (in magnitude) elements in $\boldsymbol{u}$. Then, the following holds for all $r \geq 1$:*

$$\langle \boldsymbol{u}, \boldsymbol{z}\rangle \leq \sqrt{\left\lceil \frac{|W|}{r}\right\rceil}\,\|\boldsymbol{u}_U\|\cdot\|\boldsymbol{z}_W\|.$$

Lemma 11. *Suppose that $F(\boldsymbol{x})$ is m_K-restricted strongly convex and M_K-restricted smooth for some sparsity level $K > 0$. Then for all $\eta > 0$, all vectors $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^d$ and for any Hessian matrix $\boldsymbol{H}$ of $F(\boldsymbol{x})$, we have*

$$|\langle \boldsymbol{x}, (\boldsymbol{I}-\eta\boldsymbol{H})\boldsymbol{x}'\rangle| \leq \rho\,\|\boldsymbol{x}\|\cdot\|\boldsymbol{x}'\|, \quad \textit{if } |\mathrm{supp}\,(\boldsymbol{x})\cup\mathrm{supp}\,(\boldsymbol{x}')| \leq K,$$

and

$$\|((\boldsymbol{I}-\eta\boldsymbol{H})\boldsymbol{x})_S\| \leq \rho\,\|\boldsymbol{x}\|, \quad \textit{if } |S\cup\mathrm{supp}\,(\boldsymbol{x})| \leq K,$$

where

$$\rho = \max\big\{\,|\eta m_K - 1|,\ |\eta M_K - 1|\,\big\}.$$

Proof. Since $\boldsymbol{H}$ is a Hessian matrix, we always have a decomposition $\boldsymbol{H} = \boldsymbol{A}^\top \boldsymbol{A}$ for some matrix $\boldsymbol{A}$. Denote $T = \mathrm{supp}\,(\boldsymbol{x}) \cup \mathrm{supp}\,(\boldsymbol{y})$. By simple algebra, we have

$$\begin{aligned}
\left|\langle \boldsymbol{x}, (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{x}' \rangle\right| &= \left|\langle \boldsymbol{x}, \boldsymbol{x}' \rangle - \eta \left\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}' \right\rangle\right| \\
&\overset{\zeta_1}{=} \left|\langle \boldsymbol{x}, \boldsymbol{x}' \rangle - \eta \left\langle \boldsymbol{A}_T\boldsymbol{x}, \boldsymbol{A}_T\boldsymbol{x}' \right\rangle\right| \\
&= \left|\left\langle \boldsymbol{x}, (\boldsymbol{I} - \eta \boldsymbol{A}_T^\top \boldsymbol{A}_T)\boldsymbol{x}' \right\rangle\right| \\
&\leq \left\|\boldsymbol{I} - \eta \boldsymbol{A}_T^\top \boldsymbol{A}_T\right\| \cdot \|\boldsymbol{x}\| \cdot \|\boldsymbol{x}'\| \\
&\overset{\zeta_2}{\leq} \max\left\{|\eta m_K - 1|, |\eta M_K - 1|\right\} \cdot \|\boldsymbol{x}\| \cdot \|\boldsymbol{x}'\|.
\end{aligned}$$

Here, ζ_1 follows from the fact that $\mathrm{supp}\,(\boldsymbol{x}) \cup \mathrm{supp}\,(\boldsymbol{y}) = T$ and ζ_2 holds because the RSC and RSS properties imply that the singular values of any Hessian matrix restricted on an K-sparse support set are lower and upper bounded by m_K and M_K, respectively.

For some index set S subject to $|S \cup \mathrm{supp}\,(\boldsymbol{x})| \leq K$, let $\boldsymbol{x}' = ((\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{x})_S$. We immediately obtain

$$\|\boldsymbol{x}'\|^2 = \langle \boldsymbol{x}', (\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{x} \rangle \leq \rho \|\boldsymbol{x}'\| \cdot \|\boldsymbol{x}\|,$$

indicating

$$\|((\boldsymbol{I} - \eta \boldsymbol{H})\boldsymbol{x})_S\| \leq \rho \|\boldsymbol{x}\|.$$

□

Lemma 12. *Suppose that $F(\boldsymbol{x})$ is m_K-restricted strongly convex and M_K-restricted smooth for some sparsity level $K > 0$. For all $\eta > 0$, all vectors $\boldsymbol{x}$, $\boldsymbol{x}' \in \mathbb{R}^d$ and support set T such that $|\mathrm{supp}\,(\boldsymbol{x} - \boldsymbol{x}') \cup T| \leq K$, the following holds:*

$$\|(\boldsymbol{x} - \boldsymbol{x}' - \eta \nabla F(\boldsymbol{x}) + \eta \nabla F(\boldsymbol{x}'))_T\| \leq \rho \|\boldsymbol{x} - \boldsymbol{x}'\|$$

where ρ is given in Lemma 11.

Proof. In fact, for any two vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$, there always exists a quantity $\theta \in [0, 1]$, such that

$$\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{x}') = \nabla^2 F\left(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{x}'\right)(\boldsymbol{x} - \boldsymbol{x}').$$

Let $\boldsymbol{H} = \nabla^2 F(\theta \boldsymbol{x} + (1 - \theta)\boldsymbol{x}')$. We write

$$\begin{aligned}
\|(\boldsymbol{x} - \boldsymbol{x}' - \eta \nabla F(\boldsymbol{x}) + \eta \nabla F(\boldsymbol{x}'))_T\| &= \|(\boldsymbol{x} - \boldsymbol{x}' - \eta \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}'))_T\| \\
&= \|((\boldsymbol{I} - \eta \boldsymbol{H})(\boldsymbol{x} - \boldsymbol{x}'))_T\| \\
&\leq \rho \|\boldsymbol{x} - \boldsymbol{x}'\|,
\end{aligned}$$

where the last inequality applies Lemma 11. □

Lemma 13. *Suppose that $\boldsymbol{x}$ is a k-sparse vector and let $\mathbf{b} = \boldsymbol{x} - \eta \nabla F(\boldsymbol{x})$. Let T be the support set that contains the k largest absolute values of $\mathbf{b}$. Assume that the function $F(\boldsymbol{x})$ is M_{2k}-restricted smooth, then we have the following:*

$$F(\boldsymbol{b}_T) \leq F(\boldsymbol{x}) - \frac{1 - \eta M_{2k}}{2\eta} \|\boldsymbol{b}_T - \boldsymbol{x}\|^2.$$

Proof. The RSS condition implies that

$$\begin{aligned}
F(\boldsymbol{b}_T) - F(\boldsymbol{x}) &\leq \langle \nabla F(\boldsymbol{x}), \boldsymbol{b}_T - \boldsymbol{x} \rangle + \frac{M_{2k}}{2} \|\boldsymbol{b}_T - \boldsymbol{x}\|^2 \\
&\leq -\frac{1}{2\eta} \|\boldsymbol{b}_T - \boldsymbol{x}\|^2 + \frac{M_{2k}}{2} \|\boldsymbol{b}_T - \boldsymbol{x}\|^2,
\end{aligned}$$

where the second inequality is due to the fact that

$$\begin{aligned}\|\boldsymbol{b}_T - \boldsymbol{b}\|^2 &= \|\boldsymbol{b}_T - \boldsymbol{x} + \eta \nabla F(\boldsymbol{x})\|^2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{x} + \eta \nabla F(\boldsymbol{x})\|^2 \\ &= \|\eta \nabla F(\boldsymbol{x})\|^2,\end{aligned}$$

implying

$$2\eta \langle \nabla F(\boldsymbol{x}), \boldsymbol{b}_T - \boldsymbol{x} \rangle \leq - \|\boldsymbol{b}_T - \boldsymbol{x}\|^2.$$

□

Lemma 14. *Suppose that $F(\boldsymbol{x})$ is m_K-RSC. Then for any vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ with $\|\boldsymbol{x} - \boldsymbol{x}'\|_0 \leq K$, the following holds:*

$$\|\boldsymbol{x} - \boldsymbol{x}'\| \leq \sqrt{\frac{2\max\{F(\boldsymbol{x}) - F(\boldsymbol{x}'), 0\}}{m_K}} + \frac{2\|\nabla_T F(\boldsymbol{x}')\|}{m_K}$$

where $T = \mathrm{supp}\,(\boldsymbol{x} - \boldsymbol{x}')$.

Proof. The RSC property immediately implies

$$\begin{aligned}F(\boldsymbol{x}) - F(\boldsymbol{x}') &\geq \langle \nabla F(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle + \frac{m_K}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2 \\ &\geq -\|\nabla_T F(\boldsymbol{x}')\| \cdot \|\boldsymbol{x} - \boldsymbol{x}'\| + \frac{m_K}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2.\end{aligned}$$

Discussing the sign of $F(\boldsymbol{x}) - F(\boldsymbol{x}')$ and solving the above quadratic inequality completes the proof. □

Lemma 15. *Assume that $F(\boldsymbol{x})$ is m_{k+s}-RSC and M_{2k}-RSS. Suppose that for all $t \geq 0$, $\boldsymbol{x}^t$ is k-sparse and the following holds:*

$$F(\boldsymbol{x}^{t+1}) - F(\bar{\boldsymbol{x}}) \leq \mu \left(F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}})\right) + \tau,$$

where $0 < \mu < 1$, $\tau \geq 0$ and $\bar{\boldsymbol{x}}$ is an arbitrary s-sparse signal. Then,

$$\left\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\right\| \leq \sqrt{\frac{2M}{m}}(\sqrt{\mu})^t \left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\| + \frac{3}{m}\|\nabla_{k+s} F(\bar{\boldsymbol{x}})\| + \sqrt{\frac{2\tau}{m(1-\mu)}}.$$

Proof. The RSS property implies that

$$\begin{aligned}F(\boldsymbol{x}^0) - F(\bar{\boldsymbol{x}}) &\leq \left\langle \nabla F(\bar{\boldsymbol{x}}), \boldsymbol{x}^0 - \bar{\boldsymbol{x}} \right\rangle + \frac{M}{2}\left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\|^2 \\ &\leq \frac{M}{2}\left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\|^2 + \frac{1}{2M}\|\nabla_{k+s} F(\bar{\boldsymbol{x}})\|^2 + \frac{M}{2}\left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\|^2 \\ &\leq M\left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\|^2 + \frac{1}{2M}\|\nabla_{k+s} F(\bar{\boldsymbol{x}})\|^2.\end{aligned}$$

Hence,

$$F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}}) \leq \mu^t M\left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\|^2 + \frac{1}{2M}\|\nabla_{k+s} F(\bar{\boldsymbol{x}})\|^2 + \frac{\tau}{1-\mu}.$$

By Lemma 14, we have

$$\begin{aligned}\left\|\boldsymbol{x}^t-\bar{\boldsymbol{x}}\right\| &\leq \sqrt{\frac{2}{m}}\sqrt{\mu^t M\left\|\boldsymbol{x}^0-\bar{\boldsymbol{x}}\right\|^2+\frac{\left\|\nabla_{k+s}F(\bar{\boldsymbol{x}})\right\|^2}{2M}+\frac{\tau}{1-\mu}}+\frac{2}{m}\left\|\nabla_{k+s}F(\bar{\boldsymbol{x}})\right\| \\ &\leq \sqrt{\frac{2M}{m}}(\sqrt{\mu})^t\left\|\boldsymbol{x}^0-\bar{\boldsymbol{x}}\right\|+\sqrt{\frac{1}{mM}}\left\|\nabla_{k+s}F(\bar{\boldsymbol{x}})\right\| \\ &\quad+\frac{2}{m}\left\|\nabla_{k+s}F(\bar{\boldsymbol{x}})\right\|+\sqrt{\frac{2\tau}{m(1-\mu)}} \\ &\leq \sqrt{\frac{2M}{m}}(\sqrt{\mu})^t\left\|\boldsymbol{x}^0-\bar{\boldsymbol{x}}\right\|+\frac{3}{m}\left\|\nabla_{k+s}F(\bar{\boldsymbol{x}})\right\|+\sqrt{\frac{2\tau}{m(1-\mu)}}.\end{aligned}$$

□

Lemma 16. *Let $\bar{\boldsymbol{x}} \in \mathbb{R}^d$ be an s-sparse vector supported on S. For a k-sparse vector $\boldsymbol{x}$ supported on Q with $k \geq s$, let $\boldsymbol{b} = \boldsymbol{x} - \eta\nabla F(\boldsymbol{x})$ and let $T = \mathrm{supp}\,(\boldsymbol{b}, k)$. Suppose that the function $F(\boldsymbol{x})$ is m_{2k+s}-RSC and M_{2k+s}-RSS. Then we have*

$$\left\|\bar{\boldsymbol{x}}_{S\backslash T}\right\| \leq \nu\rho\left\|\boldsymbol{x}-\bar{\boldsymbol{x}}\right\|+\nu\eta\left\|\nabla_{T\Delta S}F(\bar{\boldsymbol{x}})\right\|,$$

where $\nu = \sqrt{1+s/k}$ and ρ is given by Lemma 11.

Proof. We note the fact that the support sets $T\backslash S$ and $S\backslash T$ are disjoint. Moreover, the set $T\backslash S$ contains $|T\backslash S|$ number of top $|T|$ elements of $\boldsymbol{b}$. Hence, we have

$$\frac{1}{|T\backslash S|}\left\|\boldsymbol{b}_{T\backslash S}\right\|^2 \geq \frac{1}{|S\backslash T|}\left\|\boldsymbol{b}_{S\backslash T}\right\|^2. \tag{8}$$

That is,

$$\left\|\boldsymbol{b}_{T\backslash S}\right\| \geq \sqrt{\frac{|T\backslash S|}{|S\backslash T|}}\left\|\boldsymbol{b}_{S\backslash T}\right\| = \sqrt{\frac{k-|T\cap S|}{s-|T\cap S|}}\left\|\boldsymbol{b}_{S\backslash T}\right\| \geq \sqrt{\frac{k}{s}}\left\|\boldsymbol{b}_{S\backslash T}\right\|.$$

Note that the above holds also for $T = S$. Since $\bar{\boldsymbol{x}}$ is supported on S, the left hand side reads as

$$\left\|\boldsymbol{b}_{T\backslash S}\right\| = \left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{T\backslash S}\right\|,$$

while the right hand side reads as

$$\begin{aligned}\left\|\boldsymbol{b}_{S\backslash T}\right\| &= \left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{S\backslash T}+\bar{\boldsymbol{x}}_{S\backslash T}\right\| \\ &\geq \left\|\bar{\boldsymbol{x}}_{S\backslash T}\right\|-\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{S\backslash T}\right\|.\end{aligned}$$

Denote $\nu = \sqrt{1+s/k}$. In this way, we arrive at

$$\begin{aligned}\left\|\bar{\boldsymbol{x}}_{S\backslash T}\right\| &\leq \sqrt{\frac{s}{k}}\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{T\backslash S}\right\|+\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{S\backslash T}\right\| \\ &\leq \nu\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})\right)_{T\Delta S}\right\| \\ &\leq \nu\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})+\eta\nabla F(\bar{\boldsymbol{x}})\right)_{T\Delta S}\right\|+\nu\eta\left\|\nabla_{T\Delta S}F(\bar{\boldsymbol{x}})\right\| \\ &\leq \nu\left\|\left(\boldsymbol{x}-\bar{\boldsymbol{x}}-\eta\nabla F(\boldsymbol{x})+\eta\nabla F(\bar{\boldsymbol{x}})\right)_{T\cup Q\cup S}\right\|+\nu\eta\left\|\nabla_{T\Delta S}F(\bar{\boldsymbol{x}})\right\| \\ &\leq \nu\rho_{2k+s}\left\|\boldsymbol{x}-\bar{\boldsymbol{x}}\right\|+\nu\eta\left\|\nabla_{T\Delta S}F(\bar{\boldsymbol{x}})\right\|,\end{aligned}$$

where the second inequality follows from the fact that $ax+by \leq \sqrt{a^2+b^2}\sqrt{x^2+y^2}$ and we applied Lemma 12 for the last inequality. □

Lemma 17. *Consider the HTP algorithm with exact solutions. Assume (A1). Then*

$$\left\|\nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)\right\|^2 \geq 2m\zeta \left(F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}})\right),$$

where

$$\zeta = \frac{\left|S^{t+1}\backslash S^t\right|}{\left|S^{t+1}\backslash S^t\right| + \left|S\backslash S^t\right|}.$$

Proof. The lemma holds clearly for either $S^{t+1} = S^t$ or $F(\boldsymbol{x}^t) \leq F(\bar{\boldsymbol{x}})$. Hence, in the following we only prove the result by assuming $S^{t+1} \neq S^t$ and $F(\boldsymbol{x}^t) > F(\bar{\boldsymbol{x}})$. Due to the RSC property, we have

$$F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^t) - \left\langle \nabla F(\boldsymbol{x}^t), \bar{\boldsymbol{x}} - \boldsymbol{x}^t \right\rangle \geq \frac{m_{k+s}}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^t\right\|^2,$$

which implies

$$\begin{aligned} \left\langle \nabla F(\boldsymbol{x}^t), -\bar{\boldsymbol{x}} \right\rangle &\geq \frac{m_{k+s}}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^t\right\|^2 + F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}}) \\ &\geq \sqrt{2m_{k+s}} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^t\right\| \sqrt{F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}})}. \end{aligned}$$

By invoking Lemma 10 with $\boldsymbol{u} = \nabla F(\boldsymbol{x}^t)$ and $\boldsymbol{z} = -\bar{\boldsymbol{x}}$ therein, we have

$$\begin{aligned} \left\langle \nabla F(\boldsymbol{x}^t), -\bar{\boldsymbol{x}} \right\rangle &\leq \sqrt{\frac{|S\backslash S^t|}{|S^{t+1}\backslash S^t|} + 1} \left\|\nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)\right\| \cdot \left\|\bar{\boldsymbol{x}}_{S\backslash S^t}\right\| \\ &= \sqrt{\frac{|S\backslash S^t|}{|S^{t+1}\backslash S^t|} + 1} \left\|\nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)\right\| \cdot \left\|(\bar{\boldsymbol{x}} - \boldsymbol{x}^t)_{S\backslash S^t}\right\| \\ &\leq \sqrt{\frac{|S\backslash S^t|}{|S^{t+1}\backslash S^t|} + 1} \left\|\nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)\right\| \cdot \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^t\right\|. \end{aligned}$$

It is worth mentioning that the first inequality above holds because $\nabla F(\boldsymbol{x}^t)$ is supported on $\overline{S^t}$ and $S^{t+1}\backslash S^t$ contains the $\left|S^{t+1}\backslash S^t\right|$ number of largest (in magnitude) elements of $\nabla F(\boldsymbol{x}^t)$. Therefore, we obtain the result. □

B. Proofs for Section 2

B.1. Proof for Prop. 1

Proof. Due to the RSS property, we have

$$\begin{aligned} F(\boldsymbol{b}^{t+1}_{S^{t+1}}) - F(\boldsymbol{x}^t) &\leq \left\langle \nabla F(\boldsymbol{x}^t), \boldsymbol{b}^{t+1}_{S^{t+1}} - \boldsymbol{x}^t \right\rangle + \frac{M}{2} \left\|\boldsymbol{b}^{t+1}_{S^{t+1}} - \boldsymbol{x}^t\right\|^2 \\ &\overset{\zeta_1}{=} \left\langle \nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t), \boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t} \right\rangle + \frac{M}{2} \Big(\left\|\boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t}\right\|^2 \\ &\quad + \left\|\boldsymbol{b}^{t+1}_{S^{t+1}\cap S^t} - \boldsymbol{x}^t_{S^{t+1}\cap S^t}\right\|^2 + \left\|\boldsymbol{x}^t_{S^t\backslash S^{t+1}}\right\|^2 \Big) \\ &\overset{\zeta_2}{\leq} \left\langle \nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t), \boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t} \right\rangle + M \left\|\boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t}\right\|^2 \\ &\overset{\zeta_3}{=} -\eta(1 - \eta M) \left\|\nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)\right\|^2. \end{aligned}$$

Above, we observe that $\nabla F(\boldsymbol{x}^t)$ is supported on $\overline{S^t}$ and we simply docompose the support set $S^{t+1} \cup S^t$ into three mutually disjoint sets, and hence ζ_1 holds. To see why ζ_2 holds, we note that for any set $\Omega \subset S^t$, $\boldsymbol{b}^{t+1}_\Omega = \boldsymbol{x}^t_\Omega$. Hence, $\boldsymbol{b}^{t+1}_{S^{t+1}\cap S^t} = \boldsymbol{x}^t_{S^{t+1}\cap S^t}$. Moreover, since $\boldsymbol{x}^t_{S^t\backslash S^{t+1}} = \boldsymbol{b}^{t+1}_{S^t\backslash S^{t+1}}$ and any element in $\boldsymbol{b}^{t+1}_{S^t\backslash S^{t+1}}$ is not larger than that in $\boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t}$ (recall that S^{t+1} is obtained by hard thresholding), we have $\left\|\boldsymbol{x}^t_{S^t\backslash S^{t+1}}\right\| \leq \left\|\boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t}\right\|$ where we use the fact that $\left|S^t\backslash S^{t+1}\right| = \left|S^{t+1}\backslash S^t\right|$. Therefore, ζ_2 holds. Finally, we write $\boldsymbol{b}^{t+1}_{S^{t+1}\backslash S^t} = -\eta \nabla_{S^{t+1}\backslash S^t} F(\boldsymbol{x}^t)$ and obtain ζ_3.

Since $\boldsymbol{x}^{t+1}$ is a minimizer of $F(\boldsymbol{x})$ over the support set S^{t+1}, it immediately follows that

$$F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^t) \leq F(\boldsymbol{b}^{t+1}_{S^{t+1}}) - F(\boldsymbol{x}^t) \leq -\eta(1-\eta M) \left\| \nabla_{S^{t+1} \setminus S^t} F(\boldsymbol{x}^t) \right\|^2 .$$

Now we invoke Lemma 17 and pick $\eta \leq 1/M$,

$$F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^t) \leq \eta(\eta M - 1) \cdot \frac{2m}{1+s} \left(F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}}) \right),$$

which gives

$$F(\boldsymbol{x}^{t+1}) - F(\bar{\boldsymbol{x}}) \leq \beta \left(F(\boldsymbol{x}^t) - F(\bar{\boldsymbol{x}}) \right),$$

where $\beta = 1 - \frac{2m\eta(1-\eta M)}{1+s}$. □

B.2. Proof for Prop. 2

Proof. This is a direct result by combining Prop. 1 and Lemma 15. □

B.3. Proof for Lemma 3

Proof. Let $\boldsymbol{x}^t_* = \arg\min_{\mathrm{supp}(\boldsymbol{x}) \subset S^t} F(\boldsymbol{x})$. Since $\boldsymbol{x}^t$ and $\boldsymbol{x}^t_*$ are both supported on S^t, we apply Lemma 9 and obtain

$$\begin{aligned}
\left\| \nabla_{S^t} F(\boldsymbol{x}^t) \right\|^2 &= \left\| \nabla_{S^t} F(\boldsymbol{x}^t) - \nabla_{S^t} F(\boldsymbol{x}^t_*) \right\|^2 \\
&\leq 2M \left(F(\boldsymbol{x}^t) - F(\boldsymbol{x}^t_*) - \langle \nabla F(\boldsymbol{x}^t_*), \boldsymbol{x}^t - \boldsymbol{x}^t_* \rangle \right) \\
&\leq 2M\epsilon.
\end{aligned}$$

Above, the second inequality uses the fact that $\nabla_{S^t} F(\boldsymbol{x}^t_*) = 0$ and $F(\boldsymbol{x}^t) \leq F(\boldsymbol{x}^t_*) + \epsilon$. □

B.4. Proof for Prop. 4

Proof. We have by Lemma 16 that

$$\left\| \bar{\boldsymbol{x}}_{\overline{S^{t+1}}} \right\| \leq \sqrt{2}\rho \left\| \boldsymbol{x}^t - \bar{\boldsymbol{x}} \right\| + \frac{2}{m} \left\| \nabla_{k+s} F(\bar{\boldsymbol{x}}) \right\|,$$

where $\rho = 1 - \eta m$. On the other hand, Lemma 18 together with Lemma 3 shows that

$$\left\| \boldsymbol{x}^{t+1} - \bar{\boldsymbol{x}} \right\| \leq \kappa \left\| \bar{\boldsymbol{x}}_{\overline{S^{t+1}}} \right\| + \frac{1}{m} \left\| \nabla_k F(\bar{\boldsymbol{x}}) \right\| + \frac{1}{m} \sqrt{2M\epsilon}.$$

Therefore,

$$\left\| \boldsymbol{x}^{t+1} - \bar{\boldsymbol{x}} \right\| \leq \sqrt{2}\kappa\rho \left\| \boldsymbol{x}^t - \bar{\boldsymbol{x}} \right\| + \frac{3\kappa}{m} \left\| \nabla_{k+s} F(\bar{\boldsymbol{x}}) \right\| + \frac{\sqrt{2M\epsilon}}{m}$$

We need to ensure

$$\sqrt{2}\kappa(1 - \eta m) < 1.$$

Let $\eta = \eta'/M$ with $\eta' < 1$. Then, the above holds provided that

$$\kappa < 1 + \frac{1}{\sqrt{2}} \text{ and } \eta' > \kappa - \frac{1}{\sqrt{2}}.$$

By induction and picking proper η' to make $\sqrt{2}\kappa(1 - \eta m) < \sqrt{2}/4$, we have

$$\left\| \boldsymbol{x}^t - \bar{\boldsymbol{x}} \right\| \leq (\sqrt{2}(\kappa - \eta'))^t \left\| \boldsymbol{x}^0 - \bar{\boldsymbol{x}} \right\| + \frac{6\kappa}{m} \left\| \nabla_{k+s} F(\bar{\boldsymbol{x}}) \right\| + \frac{4\sqrt{M\epsilon}}{m}.$$

□

B.5. Proof for Prop. 5

Proof. Our proof in this part is inspired by Yuan et al. (2016). Let $\boldsymbol{x}_*^t = \arg\min_{\mathrm{supp}(\boldsymbol{x})\subset S^t} F(\boldsymbol{x})$. Then

$$\begin{aligned} F(\boldsymbol{x}^t) - F(\boldsymbol{x}^{t-1}) &\leq F(\boldsymbol{x}_*^t) - F(\boldsymbol{x}^{t-1}) + \epsilon \\ &\leq F(\boldsymbol{b}_{S^t}^t) - F(\boldsymbol{x}^{t-1}) + \epsilon \\ &\leq -\frac{1-\eta M}{2\eta}\left\|\boldsymbol{b}_{S^t}^t - \boldsymbol{x}^{t-1}\right\|^2 + \epsilon, \end{aligned}$$

where the last inequality follows from Lemma 13. Now we bound the term $\left\|\boldsymbol{b}_{S^t}^t - \boldsymbol{x}^{t-1}\right\|^2$. Note that $\boldsymbol{x}^{t-1}$ is supported on S^{t-1}. Hence,

$$\begin{aligned} \left\|\boldsymbol{b}_{S^t}^t - \boldsymbol{x}^{t-1}\right\|^2 &= \left\|\boldsymbol{x}_{S^t\cap S^{t-1}}^{t-1} - \eta\nabla_{S^t} F(\boldsymbol{x}^{t-1}) - \boldsymbol{x}^{t-1}\right\|^2 \\ &= \left\|-\boldsymbol{x}_{S^{t-1}\backslash S^t}^{t-1} - \eta\nabla_{S^t} F(\boldsymbol{x}^{t-1})\right\|^2 \\ &= \left\|\boldsymbol{x}_{S^{t-1}\backslash S^t}^{t-1}\right\|^2 + \eta^2\left\|\nabla_{S^t} F(\boldsymbol{x}^{t-1})\right\|^2 \\ &\geq \eta^2\left\|\nabla_{S^t\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2. \end{aligned}$$

We thus have

$$F(\boldsymbol{x}^t) - F(\boldsymbol{x}^{t-1}) \leq -\frac{(1-\eta M)\eta}{2}\left\|\nabla_{S^t\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \epsilon.$$

Denote $\xi = \left\|\nabla_{S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|$. We claim that

$$\left\|\nabla_{S^t\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 \geq m\left(F(\boldsymbol{x}^{t-1}) - F(\bar{\boldsymbol{x}})\right) - 2\xi^2, \tag{9}$$

which, combined with Lemma 3, immediately shows

$$F(\boldsymbol{x}^t) - F(\boldsymbol{x}^{t-1}) \leq -\frac{(1-\eta M)\eta m}{2}\left(F(\boldsymbol{x}^{t-1}) - F(\bar{\boldsymbol{x}})\right) + 2\epsilon.$$

Using Lemma 15 completes the proof.

To show (9), we consider two exhaustive cases: $\left|S^t\backslash S^{t-1}\right| \geq s$ and $\left|S^t\backslash S^{t-1}\right| < s$, and prove that (9) holds for both cases.

Case I. $\left|S^t\backslash S^{t-1}\right| \geq s$. Due to the RSC property, we have

$$\begin{aligned} &\frac{m}{2}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \\ &\leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) - \left\langle\nabla F(\boldsymbol{x}^{t-1}), \bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\rangle \\ &\leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{2}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{2m}\left\|\nabla_{S\cup S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 \\ &= F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{2}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{2m}\left\|\nabla_{S\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{2m}\left\|\nabla_{S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 \\ &= F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{2}\left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{2m}\left\|\nabla_{S\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{2m}\xi^2. \end{aligned}$$

Therefore, we get

$$\left\|\nabla_{S\backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 \geq 2m\left(F(\boldsymbol{x}^{t-1}) - F(\bar{\boldsymbol{x}})\right) - \xi^2.$$

Since S^t contains the k largest absolute values of $\boldsymbol{b}^t$, and $\left|S^t\backslash S^{t-1}\right| \geq s \geq \left|S\backslash S^{t-1}\right|$, we have

$$\left\|\boldsymbol{b}_{S^t\backslash S^{t-1}}^t\right\|^2 \geq \left\|\boldsymbol{b}_{S\backslash S^{t-1}}^t\right\|^2,$$

which immediately implies (9) by noting the fact that $\boldsymbol{b}^t_{S^t \backslash S^{t-1}} = -\eta \nabla_{S^t \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})$ and $\boldsymbol{b}^t_{S \backslash S^{t-1}} = -\eta \nabla_{S \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})$.

Case II. $\left|S^t \backslash S^{t-1}\right| < s$. Again, we use the RSC property to obtain

$$\begin{aligned}
&\frac{m}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \\
&\leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) - \left\langle \nabla F(\boldsymbol{x}^{t-1}), \bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1} \right\rangle \\
&\leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{4} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{m} \left\|\nabla_{S \cup S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 \\
&= F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{4} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{m} \left\|\nabla_{S \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{m}\xi^2 \\
&= F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{4} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{m} \left\|\nabla_{S \backslash (S^t \cup S^{t-1})} F(\boldsymbol{x}^{t-1})\right\|^2 \\
&\quad + \frac{1}{m} \left\|\nabla_{(S^t \backslash S^{t-1}) \cap S} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{m}\xi^2 \\
&\leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{4} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{1}{m} \left\|\nabla_{S \backslash (S^t \cup S^{t-1})} F(\boldsymbol{x}^{t-1})\right\|^2 \\
&\quad + \frac{1}{m} \left\|\nabla_{S^t \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{m}\xi^2.
\end{aligned} \tag{10}$$

We consider the term $\left\|\nabla_{S \backslash (S^t \cup S^{t-1})} F(\boldsymbol{x}^{t-1})\right\|^2$ above. Actually, we have

$$\boldsymbol{b}^t_{S \backslash (S^t \cup S^{t-1})} = -\eta \nabla_{S \backslash (S^t \cup S^{t-1})} F(\boldsymbol{x}^{t-1}).$$

Since S^t contains the k largest absolute values of $\boldsymbol{b}^t$, we know that any component in $\boldsymbol{b}^t_\Omega$ is not larger than that in $\boldsymbol{b}^t_{S^t}$ subject to $\Omega \cap S^t = \emptyset$. In particular,

$$\frac{\left\|\boldsymbol{b}^t_{S \backslash (S^t \cup S^{t-1})}\right\|^2}{\left|S \backslash (S^t \cup S^{t-1})\right|} \leq \frac{\left\|\boldsymbol{b}^t_{(S^t \cap S^{t-1}) \backslash S}\right\|^2}{\left|(S^t \cap S^{t-1}) \backslash S\right|}.$$

Note that $\left|S^t \backslash S^{t-1}\right| < s$ implies $\left|(S^t \cap S^{t-1}) \backslash S\right| \geq k - 2s$. Therefore,

$$\begin{aligned}
&\eta^2 \left\|\nabla_{S \backslash (S^t \cup S^{t-1})} F(\boldsymbol{x}^{t-1})\right\|^2 \\
&\leq \frac{s}{k-2s} \left\|\boldsymbol{x}^{t-1}_{(S^t \cap S^{t-1}) \backslash S} - \eta \nabla_{(S^t \cap S^{t-1}) \backslash S} F(\boldsymbol{x}^{t-1})\right\|^2 \\
&\leq \frac{2s}{k-2s} \left\|\boldsymbol{x}^{t-1}_{(S^t \cap S^{t-1}) \backslash S}\right\|^2 + \frac{2s\eta^2}{k-2s}\xi^2 \\
&= \frac{2s}{k-2s} \left\|(\boldsymbol{x}^{t-1} - \bar{\boldsymbol{x}})_{(S^t \cap S^{t-1}) \backslash S}\right\|^2 + \frac{2s\eta^2}{k-2s}\xi^2 \\
&\leq \frac{2s}{k-2s} \left\|\boldsymbol{x}^{t-1} - \bar{\boldsymbol{x}}\right\|^2 + \frac{2s\eta^2}{k-2s}\xi^2.
\end{aligned}$$

Plugging the above into (10), we obtain

$$\begin{aligned}
\frac{m}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{4} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 + \frac{2s}{(k-2s)\eta^2 m} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \\
+ \frac{1}{m} \left\|\nabla_{S^t \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \frac{1}{m} \left(\frac{2s}{k-2s} + 1\right)\xi^2.
\end{aligned}$$

Picking $k \geq 2s + \frac{8s}{\eta^2 m^2}$ gives

$$\begin{aligned}
\frac{m}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \leq F(\bar{\boldsymbol{x}}) - F(\boldsymbol{x}^{t-1}) + \frac{m}{2} \left\|\bar{\boldsymbol{x}} - \boldsymbol{x}^{t-1}\right\|^2 \\
+ \frac{1}{m} \left\|\nabla_{S^t \backslash S^{t-1}} F(\boldsymbol{x}^{t-1})\right\|^2 + \left(\frac{\eta^2 m}{4} + \frac{1}{m}\right)\xi^2.
\end{aligned}$$

Since $\eta < 1/M$, $\frac{\eta^2 m^2}{4} + 1 < 2$. Therefore, by re-arranging the above inequality, we prove the claim (9). □

C. Proofs for Section 3

The following result holds for all $F(\boldsymbol{x})$.

Lemma 18. *Assume (A1) and (A2). For any k-sparse vector $\boldsymbol{x}$ and s-sparse vector $\bar{\boldsymbol{x}}$, we have*

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \kappa \|\bar{\boldsymbol{x}}_{\overline{T}}\| + \frac{1}{m} \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\|,$$

where T is the support set of $\boldsymbol{x}$.

Proof.

$$\begin{aligned}
\|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\|^2 &= \langle \boldsymbol{x} - \bar{\boldsymbol{x}} - \tau \nabla F(\boldsymbol{x}) + \tau \nabla F(\bar{\boldsymbol{x}}), (\boldsymbol{x} - \bar{\boldsymbol{x}})_T \rangle + \tau \langle \nabla F(\boldsymbol{x}) - \nabla F(\bar{\boldsymbol{x}}), (\boldsymbol{x} - \bar{\boldsymbol{x}})_T \rangle \\
&\leq \|(\boldsymbol{x} - \bar{\boldsymbol{x}} - \tau \nabla F(\boldsymbol{x}) + \tau \nabla F(\bar{\boldsymbol{x}}))_T\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| + \tau \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| \\
&\leq \|\boldsymbol{x} - \bar{\boldsymbol{x}} - \tau \nabla_{T \cup S} F(\boldsymbol{x}) + \tau \nabla_{T \cup S} F(\bar{\boldsymbol{x}})\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| + \tau \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| \\
&\leq \rho \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| + \tau \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\| \cdot \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\|.
\end{aligned}$$

Dividing both sides by $\|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\|$ gives

$$\|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| \leq \rho \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + \tau \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\|.$$

On the other hand,

$$\begin{aligned}
\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| &\leq \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_T\| + \|(\boldsymbol{x} - \bar{\boldsymbol{x}})_{\overline{T}}\| \\
&\leq \rho \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| + \tau \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\| + \|\bar{\boldsymbol{x}}_{\overline{T}}\|.
\end{aligned}$$

Hence, we have

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \frac{1}{1-\rho} \|\bar{\boldsymbol{x}}_{\overline{T}}\| + \frac{\tau}{1-\rho} \|\nabla_T F(\boldsymbol{x}) - \nabla_T F(\bar{\boldsymbol{x}})\|.$$

Picking $\tau = 1/M$ completes the proof. □

In view of the exact (HTP3), we have

$$\left\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\right\| \leq \kappa \left\|\bar{\boldsymbol{x}}_{\overline{S^t}}\right\| + \frac{1}{m} \|\nabla_k F(\bar{\boldsymbol{x}})\|. \tag{11}$$

Now we present the crucial lemma. It is inspired by Bouchot et al. (2016) but we show a more general result.

Lemma 19. *Consider the HTP algorithm. Assume (A1) and (A2). Further assume that the sequence of $\{\boldsymbol{x}^t\}_{t \geq 0}$ satisfies*

$$\begin{aligned}
\left\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\right\| &\leq \alpha \cdot \beta^t \left\|\boldsymbol{x}^0 - \bar{\boldsymbol{x}}\right\| + \phi, \\
\left\|\boldsymbol{x}^t - \bar{\boldsymbol{x}}\right\| &\leq \gamma \left\|\bar{\boldsymbol{x}}_{\overline{S^t}}\right\| + \psi,
\end{aligned}$$

for positive α, ϕ, γ, ψ and $0 < \beta < 1$. Suppose that at the n-th iteration ($n \geq 0$), S^n contains the indices of top p (in magnitude) elements of $\bar{\boldsymbol{x}}$. Then, for any integer $1 \leq q \leq s - p$, there exists an integer $r \geq 1$ determined by

$$\sqrt{2} \left|\bar{\boldsymbol{x}}_{p+q}\right| > \alpha\gamma \cdot \beta^{r-1} \left\|\bar{\boldsymbol{x}}_{\{p+1,\dots,s\}}\right\| + \theta$$

where

$$\theta = \alpha\psi + \phi + \frac{1}{m} \|\nabla_2 F(\bar{\boldsymbol{x}})\|,$$

such that S^{n+r} contains the indices of top $p+q$ elements of $\bar{\boldsymbol{x}}$ provided that $\theta \leq \sqrt{2}\lambda \bar{\boldsymbol{x}}_{\min}$ for some $\lambda \in (0, 1)$.

Proof. Without loss of generality, we presume that the elements in $\bar{\boldsymbol{x}}$ are in descending order by their magnitude, i.e., $|\bar{\boldsymbol{x}}_1| \geq |\bar{\boldsymbol{x}}_2| \geq \cdots \geq |\bar{\boldsymbol{x}}_s|$. We aim at deriving a condition under which $[p+q] \subset S^{n+r}$. To this end, it suffices to enforce

$$\min_{j \in [p+q]} \left|\boldsymbol{b}_j^{n+r}\right| > \max_{i \in \overline{S}} \left|\boldsymbol{b}_i^{n+r}\right|. \tag{12}$$

On one hand, for any $j \in [p+q]$,

$$\begin{aligned}
\left|\boldsymbol{b}_j^{n+r}\right| &= \left|\left(\boldsymbol{x}^{n+r-1} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right| \\
&\geq |\bar{\boldsymbol{x}}_j| - \left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right| \\
&\geq |\bar{\boldsymbol{x}}_{p+q}| - \left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right|.
\end{aligned}$$

On the other hand, for all $i \in \overline{S}$,

$$\left|\boldsymbol{b}_i^{n+r}\right| = \left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_i\right|.$$

Hence, we know that to guarantee (12), it suffices to ensure for all $j \in [p+q]$ and $i \in \overline{S}$ that

$$|\bar{\boldsymbol{x}}_{p+q}| > \left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right| + \left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_i\right|.$$

Note that the right-hand side is upper bounded as follows:

$$\begin{aligned}
&\frac{1}{\sqrt{2}}\left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right| + \frac{1}{\sqrt{2}}\left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_i\right| \\
&\leq \left\|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_{\{j,i\}}\right\| \\
&\leq \left\|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1}) + \eta \nabla F(\bar{\boldsymbol{x}})\right)_{\{j,i\}}\right\| + \eta \left\|\nabla_{\{j,i\}} F(\bar{\boldsymbol{x}})\right\| \\
&\leq \rho \left\|\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}}\right\| + \eta \left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\| \\
&\leq \rho\alpha \cdot \beta^{r-1} \left\|\boldsymbol{x}^n - \bar{\boldsymbol{x}}\right\| + \rho\phi + \eta \left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\|.
\end{aligned}$$

Moreover,

$$\left\|\boldsymbol{x}^n - \bar{\boldsymbol{x}}\right\| \leq \gamma \left\|\bar{\boldsymbol{x}}_{\overline{S^n}}\right\| + \psi \leq \gamma \left\|\bar{\boldsymbol{x}}_{\overline{[p]}}\right\| + \psi = \gamma \left\|\bar{\boldsymbol{x}}_{\{p+1,\dots,s\}}\right\| + \psi.$$

Put all together, we have

$$\begin{aligned}
&\frac{1}{\sqrt{2}}\left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_j\right| + \frac{1}{\sqrt{2}}\left|\left(\boldsymbol{x}^{n+r-1} - \bar{\boldsymbol{x}} - \eta \nabla F(\boldsymbol{x}^{n+r-1})\right)_i\right| \\
&\leq \rho\alpha\gamma \cdot \beta^{r-1} \left\|\bar{\boldsymbol{x}}_{\{p+1,\dots,s\}}\right\| + \rho\alpha\psi + \rho\phi + \eta \left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\| \\
&\leq \alpha\gamma \cdot \beta^{r-1} \left\|\bar{\boldsymbol{x}}_{\{p+1,\dots,s\}}\right\| + \alpha\psi + \phi + \frac{1}{m} \left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\|.
\end{aligned}$$

Therefore, when

$$\sqrt{2}\,|\bar{\boldsymbol{x}}_{p+q}| > \alpha\gamma \cdot \beta^{r-1} \left\|\bar{\boldsymbol{x}}_{\{p+1,\dots,s\}}\right\| + \alpha\psi + \phi + \frac{1}{m} \left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\|,$$

we always have (12). Note that the above holds as far as $\alpha\psi + \phi + \frac{1}{m}\left\|\nabla_2 F(\bar{\boldsymbol{x}})\right\|$ is strictly smaller than $\sqrt{2}\,|\bar{\boldsymbol{x}}_s|$. ☐

With Lemma 19, we show the following general theorem.

Theorem 20. *Assume same conditions as in Lemma 19. Then HTP successfully identifies the support of $\bar{\boldsymbol{x}}$ using $\left(\frac{\log 2}{2\log(1/\beta)} + \frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2\right) s$ number of iterations.*

Proof. Without loss of generality, we presume that the elements in $\bar{\boldsymbol{x}}$ are in descending order by their magnitude, i.e., $|\bar{\boldsymbol{x}}_1| \geq |\bar{\boldsymbol{x}}_2| \geq \cdots \geq |\bar{\boldsymbol{x}}_s|$. We partition the support set $[s]$ into K folds $S_1, S_2, \dots, S_K$, where each S_i is defined as follows:

$$S_i = \{s_{i-1}+1, \dots, s_i\},\ \forall\, 1 \leq i \leq K.$$

Here, $s_0 = 0$ and for all $1 \leq i \leq K$, the quantity s_i is inductively given by

$$s_i = \max\left\{q:\ s_{i-1}+1 \leq q \leq s \text{ and } |\bar{\boldsymbol{x}}_q| > \frac{1}{\sqrt{2}}\left|\bar{\boldsymbol{x}}_{s_{i-1}+1}\right|\right\}.$$

In this way, we note that for any two index sets S_i and S_j, $S_i \cap S_j = \emptyset$ if $i \neq j$. We also know by the definition of s_i that

$$|\bar{\boldsymbol{x}}_{s_i+1}| \leq \frac{1}{\sqrt{2}}\left|\bar{\boldsymbol{x}}_{s_{i-1}+1}\right|,\ \forall\, 1 \leq i \leq K-1. \tag{13}$$

Now we show that after a finite number of iterations, say n, the union of the S_i's is contained in S^n. To this end, we prove that for all $0 \leq i \leq K$,

$$\bigcup_{t=0}^{i} S_t \subset S^{n_0+n_1+\cdots+n_i} \tag{14}$$

for some n_i's given below.

We pick $n_0 = 0$ and it is easy to verify that $S_0 \subset S^0$. Now suppose that (14) holds for $i-1$. That is, the index set of the top s_{i-1} elements of $\bar{\boldsymbol{x}}$ is contained in $S^{n_0+\cdots+n_{i-1}}$. Due to Lemma 19, (14) holds for i as long as n_i satisfies

$$\sqrt{2}\,|\bar{\boldsymbol{x}}_{s_i}| > \alpha\gamma\cdot\beta^{n_i-1}\left\|\bar{\boldsymbol{x}}_{\{s_{i-1}+1,\dots,s\}}\right\| + \theta. \tag{15}$$

Note that

$$\begin{aligned}
\left\|\bar{\boldsymbol{x}}_{\{s_{i-1}+1,\dots,s\}}\right\|^2 &= \left\|\bar{\boldsymbol{x}}_{S_i}\right\|^2 + \cdots + \left\|\bar{\boldsymbol{x}}_{S_K}\right\|^2 \\
&\leq (\bar{\boldsymbol{x}}_{s_{i-1}+1})^2\,|S_i| + \cdots + (\bar{\boldsymbol{x}}_{s_{r-1}+1})^2\,|S_K| \\
&\leq (\bar{\boldsymbol{x}}_{s_{i-1}+1})^2\left(|S_i| + 2^{-1}\,|S_{i+1}| + \cdots + 2^{i-K}\,|S_K|\right) \\
&< 2(\bar{\boldsymbol{x}}_{s_i})^2\left(|S_i| + 2^{-1}\,|S_{i+1}| + \cdots + 2^{i-K}\,|S_K|\right),
\end{aligned}$$

where the second inequality follows from (13) and the last inequality follows from the definition of q_i. Denote for simplicity

$$T_i := |S_i| + 2^{-1}\,|S_{i+1}| + \cdots + 2^{i-K}\,|S_K|\,.$$

As we assume $\theta \leq \sqrt{2}\lambda\bar{\boldsymbol{x}}_{\min}$, we get

$$\alpha\gamma\cdot\beta^{n_i-1}\left\|\bar{\boldsymbol{x}}_{\{s_{i-1}+1,\dots,s\}}\right\| + \theta < \sqrt{2}\alpha\gamma\,|\bar{\boldsymbol{x}}_{s_i}|\,\beta^{n_i-1}\sqrt{T_i} + \sqrt{2}\lambda\,|\bar{\boldsymbol{x}}_{s_i}|\,.$$

Picking

$$n_i = \log_{1/\beta}\frac{\alpha\gamma\sqrt{T_i}}{1-\lambda} + 2$$

guarantees (15). It remains to calculate the total number of iterations. In fact, we have

$$\begin{aligned}
n &= n_0 + n_1 + \dots n_K \\
&= \frac{1}{2\log(1/\beta)} \sum_{i=1}^{K} \log T_i + K \cdot \frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2K \\
&\overset{\zeta_1}{\leq} \frac{K}{2\log(1/\beta)} \log\left(\frac{1}{K}\sum_{i=1}^{K} T_i\right) + \left(\frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2\right) K \\
&\overset{\zeta_2}{\leq} \frac{K}{2\log(1/\beta)} \log\left(\frac{2}{K}\sum_{i=1}^{K} |S_i|\right) + \left(\frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2\right) K \\
&= \frac{K}{2\log(1/\beta)} \log\frac{2s}{K} + \left(\frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2\right) K \\
&\overset{\zeta_3}{\leq} \left(\frac{\log 2}{2\log(1/\beta)} + \frac{\log(\alpha\gamma/(1-\lambda))}{\log(1/\beta)} + 2\right) s.
\end{aligned}$$

Above, ζ_1 immediately follows by observing that the logarithmic function is concave. ζ_2 uses the fact that after rearrangement, the coefficient of $|S_i|$ is $\sum_{j=0}^{i-1} 2^{-j}$ which is always smaller than 2. Finally, since the function $r\log(2s/r)$ is monotonically increasing with respect to r and $1 \leq r \leq s$, ζ_3 follows. □

Combining this theorem, Lemma 19 and specific results in Prop. 2, Prop. 4 and Prop. 5 gives the main theorems in Section 3.