Appendix

A Preliminary

We introduce some notations and Lemmas that appear in earlier works [Nutini et al., 2015, Karimireddy et al., 2019].

We say a gradient step is good if the post-processing step in Eq. (3) is not triggered i.e., \( x_{i}^{t+1}x_{i}^{t} \geq 0 \), otherwise we call this step a bad step. We denote the set of good steps until the \( t \)-th iteration as \( G_{t} \), since a bad step always follows a good step, it is easy to verify that

\[
|G_{t}| \leq \left\lceil \frac{t}{2} \right\rceil .
\]  

(1)

Recall the selection rule in section 3:

**Selection rule 1 (GS-s rule).** Select \( i \in \arg \max_{j} Q_{j}(x^{t}) \) where

\[
Q_{i}(x) = \min_{s \in \partial g_{i}} |\nabla f(x) + s|.
\]  

(2)

**Lemma 2 ([Karimireddy et al., 2019]).** Assume \( f(\cdot) \) is \( \mu_{1} \) strongly convex with respect to 1-norm, then the iterates generated from Algorithm 1 with GS-s rule (selection rule 2) satisfy

\[
F(x^{t}) - F(x^{\ast}) \leq \left( 1 - \frac{\mu_{1}}{L} \right)^{\frac{1}{2}} (F(0) - F(x^{\ast})).
\]

The above lemma is from [Karimireddy et al., 2019].

**Lemma 3 ([Karimireddy et al., 2019]).** Consider \( g(\cdot) \) to be \( \ell_{1} \) regularization or non-negative constraint. Then if the \( t \)-th iteration is a good step, we have

\[
F(x^{t+1}) \leq F(x^{t}) - \frac{1}{2L} \max_{i \in [d]} Q_{i}(x^{t})^{2},
\]  

(3)

where \( Q_{i}(\cdot) \) is defined in the GS-s rule (selection rule 3).

B Proof of Theorem Sketch 2

**Proof.** Let \( W = \{w_1, w_2, ..., w_k\} \) s.t. \( w_1 < w_2 < ... < w_k \in \mathbb{N} \), we define new functions \( h(\cdot) : \mathbb{R}^{k} \to \mathbb{R} \), \( h(y) = f(\sum_{i=1}^{k} y_{i}e_{w_{i}}) \) and \( H(y) := h(y) + \sum_{i=1}^{k} g_{w_{i}}(y_{i}) \).
First, we show that \( h(y + \alpha e_i) \) is also \( L \)-smooth \( \forall i \in [k] \).

For any \( i \in [k], y \in \mathbb{R}^k \),

\[
\begin{align*}
  h(y + \alpha e_i) &= f(\sum_{j=1}^{k} y_j e_w_j + \alpha e_{w_i}) \\
  &\leq f(\sum_{j=1}^{k} y_j e_w_j) + \alpha \nabla_{w_i} f(\sum_{j=1}^{k} y_j e_{w_j}) + \frac{L}{2} \alpha^2 \\
  &= h(y) + \alpha \nabla_i h(y) + \frac{L}{2} \alpha^2
\end{align*}
\]  

(4)

Second, we show that we can get the same iterates if we run GCD on \( F(x) \) or \( H(y) \), that is, we want to show that \( x^t = \sum_{i=1}^{k} e_w y_t^i \ \forall t \geq 0 \). We prove by induction:

When \( t = 0 \), obviously we have \( x_0 = \sum_{i=1}^{k} e_w y_0^i = 0 \).

Suppose that \( x^t = \sum_{i=1}^{k} e_w y_t^i, \ i = \text{arg max}_j Q_j(x^t), \ \tilde{i} = \text{arg max}_j Q_j(y^t) \) and \( i = w_m \), we can show that \( \tilde{i} = m \):

Note that

\[
Q_i(x^t) = \min_{s \in \partial g_i} |\nabla_i f(x^t) + s|
\]

\[= \min_{s \in \partial g_m} |\nabla_m h(x) + s|
\]

\[= Q_m(y^t),
\]

(5)

Thus, it is easy to see that \( \tilde{i} = m \).

\[
x^{t+\frac{1}{2}}_i = x^t_i - \frac{1}{L} \nabla f_i(x^t) = y^t_m = \frac{1}{L} \nabla h_m(y^t) = y^{t+\frac{1}{2}}_m.
\]

Note that \( g_i(\cdot) = g_{w_m}(\cdot) \), thus we further have

\[
x^{t+1}_i = \text{prox}_{\frac{1}{L} g_i} \left[ x^{t+\frac{1}{2}}_i \right] = \text{prox}_{\frac{1}{L} g_{w_m}} \left[ y^{t+\frac{1}{2}}_m \right] = y^{t+1}_m.
\]

Thus we have \( x^t = \sum_{i=1}^{k} e_w y_t^i \ \forall t = 0, 1, 2, ... \)

Plug \( H(\cdot) \) into Lemma 2 and using the above result, we can get

\[
F(x^t) - F(x^*) = H(y^t) - H(y^*) \leq \left( 1 - \frac{\tilde{\mu}_1}{L} \right)^{\lceil \frac{t}{2} \rceil} (H(0) - H(y^*)) = \left( 1 - \frac{\tilde{\mu}_1}{L} \right)^{\lceil \frac{t}{2} \rceil} (F(0) - F(x^*))
\]

where \( \tilde{\mu}_1 \) is the 1-norm strongly convex constant for the \( k \)-dimensional small problem \( H(\cdot) \), since \( H \) is also \( \mu_2 \) strongly convex, we can easily verify that \( \max\{\mu_2/k, \mu_1\} \leq \tilde{\mu}_1 \leq \mu_2 \), which completes the proof.

\[\square\]

C Proof of Lemma 3

Proof. If \( i \) is not select by Algorithm 1 at the \( t \)-th iteration, then \( x^{t+1}_i = 0 \) trivially remains 0.
If $i$ is selected at the $t$-th iteration, by assuming $|\nabla_i f(x^t) - \nabla_i f(x^*)| \leq \delta_i$, we know that

$$-\delta_i + \nabla_i f(x^*) \leq \nabla_i f(x^t) \leq \delta_i + \nabla_i f(x^*)$$

\[\Rightarrow -u_i \leq \nabla_i f(x^t) \leq -l_i,\]  

where (i) follows directly from the definition of $\delta_i := \min \{-\nabla_i f(x^*) - l_i, u_i + \nabla_i f(x^*)\}$.

Then we show that $\text{prox}_{\frac{1}{L_i} \nabla_i f(x^t)}(0 - \frac{1}{L_i} \nabla_i f(x^t)) = 0$:

$$\text{prox}_{\frac{1}{L_i} \nabla_i f(x^t)} \left(0 - \frac{1}{L_i} \nabla_i f(x^t)\right) = \arg\min_y \left\{ \frac{1}{2} \left(y - \left(-\frac{1}{L_i} \nabla f_i(x^t)\right)\right)^2 + \frac{1}{L_i} g_i(y) \right\}$$

This minimization problem is strongly convex and thus has a unique solution satisfies:

$$0 \in y + \frac{1}{L_i} \nabla_i f(x^t) + \frac{1}{L_i} \partial g_i(y)$$

By knowing $-u_i \leq \nabla_i f(x^t) \leq -l_i$ from Eq. 6 and $\text{int} \partial g_i(0) = (l_i, u_i)$ by the definition of $l_i$ and $u_i$. We can easily conclude that $y = 0$ satisfies Eq. 8 and therefore

$$x_i^{t+1} = \text{prox}_{\frac{1}{L_i} \nabla_i f(x^t)} \left(0 - \frac{1}{L_i} \nabla_i f(x^t)\right) = 0.$$  

\[\square\]

### D Proof of Theorem 4

**Proof.** Let $t \leq d - \tau$ and recall the definition of good steps until the $t$-th iteration from section A in Appendix: $|G_t| = \{i_1, i_2, \ldots, i_k\}$, where $k \geq \left\lceil \frac{1}{2} \right\rceil$.

At iteration $i_m, m \in [k], x^{i_m}$ is guaranteed to be $m - 1$-sparse, by assuming $f(\cdot)$ is $\mu_1^{(\tau + m - 1)}$ strongly convex w.r.t. 1-norm and $\tau + m - 1$-sparse vectors, we know that $F(\cdot)$ is also $\mu_1$ strongly convex w.r.t. 1-norm and $\tau + m - 1$-sparse vectors. Thus $\forall y \in \mathbb{R}^d$ that is $\tau$-sparse, $|\text{supp}(y) \cup \text{supp}(x^{i_m})| \leq \tau + m - 1$ and by the definition of $\mu_1^{(\tau + m - 1)}$, we have

$$F(y) \geq F(x^{i_m}) + \langle \partial F(x^{i_m}), y - x^{i_m} \rangle + \frac{\mu_1^{(\tau + m - 1)}}{2} \|y - x^{i_m}\|_1^2,$$

with a little bit abuse of notation, here $\partial F(x^t)$ stands for any vector in the subdifferential of $F(x^t)$. Taking minimum on both side of Eq. 9 w.r.t. $y$ that is $\tau$ sparse,

$$F(x^*) \geq F(x^{i_m}) - \sup_{\|y\|_0 \leq \tau} \left(\langle -\partial F(x^{i_m}), y - x^{i_m} \rangle - \frac{\mu_1^{(\tau + m - 1)}}{2} \|y - x^{i_m}\|_1^2\right)$$

$$\geq F(x^{i_m}) - \sup_{y \in \mathbb{R}^d} \left(\langle -\partial F(x^{i_m}), y - x^{i_m} \rangle - \frac{\mu_1^{(\tau + m - 1)}}{2} \|y - x^{i_m}\|_1^2\right)$$

\[\Rightarrow F(x^{i_m}) - \frac{\mu_1^{(\tau + m - 1)}}{2} \|\cdot\|_1^n (-\partial F(x^{i_m}))\]  

\[\Rightarrow F(x^{i_m}) - \frac{1}{2\mu_1^{(\tau + m - 1)}} \|\partial F(x^{i_m})\|_\infty^2.

3
where (i) is from the definition of conjugate function, and (ii) is from the fact that \( \left( \frac{1}{2} \| \cdot \|_1 \right)^* = \frac{1}{2} \| \cdot \|_\infty \) Boyd and Vandenberghe, 2004.

More specifically,

\[
F(x^*) \geq F(x^m) - \frac{1}{2\mu_1^{(r+m-1)}} \| \nabla f(x^m) + u \|_\infty^2 \quad \forall u \in \partial g(x^m).
\]

By the definition of \( Q_i(\cdot) \) in the GS-s rule (selection rule 2), we further have

\[
F(x^*) \geq F(x^m) - \frac{1}{2\mu_1^{(r+m-1)}} \max_{i \in [d]} Q_i(x^m)^2.
\]

Recall Lemma 3, we have

\[
F(x^{m+1}) \geq F(x^m) - \frac{1}{2L} \max_{i \in [d]} Q_i(x^m)^2.
\]

Plug the above equation into Eq. (10)

\[
F(x^*) \geq F(x^m) - \frac{L}{\mu_1^{(r+m-1)}} (F(x^{m+1}) - F(x^m))
\]

\[
\Rightarrow F(x^{m+1}) - F^* \leq \left( 1 - \frac{\mu_1^{(r+m)}}{L} \right) (F(x^m) - F^*).
\]

By applying the above inequality recursively, we get

\[
F(x^t) - F^* \leq \prod_{m=1}^{k} \left( 1 - \frac{\mu_1^{(r+m-1)}}{L} \right) (F(0) - F^*)
\]

\[
\leq \prod_{i=1}^{\left\lceil \frac{t}{2} \right\rceil} \left( 1 - \frac{\mu_1^{(r+i-1)}}{L} \right) (F(0) - F^*),
\]

which completes the proof.

\[\Box\]

**E Proof of Theorem 8**

**Proof.** This proof is essentially the same as Theorem 4, the difference is that, by the definition of the \( \Delta \)-GS-s rule (selection rule 7), the Lemma 3 becomes

\[
F(x^{t+1}) - F(x^t) \leq -\frac{\Delta}{2L} \max_{i \in [d]} Q_i(x^t)^2
\]

at each good step \( t \).

Knowing that \( \text{supp}(x^t) \subset W_\Delta \), we have \( |\text{supp}(x^t) \cup \text{supp}(x^t)| \leq |W_\Delta| \forall t > 0 \). Then we can incorporate the new Lemma into the analysis of Theorem 4 and get

\[
F(x^t) - F^* \leq \left( 1 - \frac{\Delta \mu_1^{(W_\Delta)}}{L} \right)^{\left\lceil \frac{t}{2} \right\rceil} (F(0) - F^*)
\]

\[
\leq \left( 1 - \frac{\Delta \mu_2}{|W_\Delta|L} \right)^{\left\lceil \frac{t}{2} \right\rceil} (F(0) - F^*).
\]

\[\Box\]
F Proof of Theorem 9

Proof.

Clarify some notations

Given $\Delta > 0$, we sort $W_\Delta = \{i_1, i_2, \ldots, i_m\}$ by the number of iteration when they first enter the working set $W_\Delta$ i.e., $i_1$ is the first coordinate being selected and $i_2$ is the second coordinate to be included in $W_\Delta$, etc.

We denote the $t$-th iterate from the $\Delta$-GCD algorithm as $x^t$ and the $t$-th iterate from the totally corrective greedy algorithm (TCGA) as $\tilde{x}^t$. $W^2 = \{i_1, i_2, \ldots, i_k\}$, its elements is also sorted by the time when they enter the working set.

A claim:

First, we show that $\forall j \leq k$, there $\exists \epsilon_j > 0$ such that $\forall \Delta < \epsilon_j$, the first $j$ elements in $W_\Delta$ is the same as the first $j$ elements in $W^2$.

We prove this claim by induction, when $j = 1$, $\forall \Delta \leq 1$, $\Delta$-GCD and the TCGA both select the coordinate $\arg \max_{i \in [d]} Q_i(0)$ at the first iteration, thus the claim is true in this base case.

Assuming that the claim is true with some $j > 0$, then for $j + 1$:

By the continuity of $Q_i(\cdot)$, we know that there $\exists \epsilon'$ such that $\forall \|x - \tilde{x}^j\| \leq \epsilon'$, $\arg \max_{i \in [d]} Q_i(x) = \tilde{i}_{j+1}$.

By the uniqueness (recall that $F(\cdot)$ is strongly convex) of $\tilde{x}^j$:

$$\tilde{x}^j := \arg \min_{\text{supp}(x) \subseteq W_j} f(x) + g(x)$$

and the optimality condition, we also know that there $\exists \delta > 0$ such that $\forall x \in \mathbb{R}^d$ satisfy $\text{supp}(x) \subseteq W_j$ and $\max_{i \in W_j} Q_i(x) \leq \delta$, we have $\|x - \tilde{x}^j\| \leq \epsilon'$.

Denote $Q_i(x^t)$ (recall $x^t$ is generated from $\Delta$-GCD) is bounded by some constant $B \forall t > 0$.

Then, by setting $\Delta \leq (\min\{\epsilon_j, \delta/B\})^2$, when $i_{j+1}$ first enter $W_\Delta$ at some iteration $t$, we have

$$\arg \max_{i \in W_j} Q_i(x^t) \leq \sqrt{\Delta} \arg \max_{i \in [d]} Q_i(x^t) \leq \frac{\delta}{B} = \delta,$$

also by the induction assumption, we know that $\text{supp}(x^t) \subseteq W_j$. Putting these two conditions together, we get $\|x^t - \tilde{x}^j\| \leq \epsilon'$ and thus $\arg \max_{i \in [d]} Q_i(x^t) = \tilde{i}_{j+1}$, which implies that $i_{j+1} = \tilde{i}_{j+1}$. And this complete the proof of this claim.

Back to the proof:

Following the claim, we know that there $\exists \epsilon_k > 0$ such that for $\forall \Delta < \epsilon_k$, the first $k$ elements in $W_\Delta$ is just $W^2$.

By the nondegeneracy assumption i.e., $\delta_j > 0 \forall x^*_i = 0$ and continuity of $Q_i(\cdot), \nabla f(\cdot)$, we know that there $\exists \epsilon'' > 0$ such that $\forall \|x - x^*\| < \epsilon''$ (note that $\tilde{x}^k = x^*$), $|\nabla_i f(x) - \nabla_i f(x^*)| \leq \delta_i \forall x^*_i = 0$ and this further implies $Q_i(x) = 0 \forall i \notin W^2$ (note that $\text{supp}(x^*) \in W^2$).

Again, there exist $\delta'' > 0$ such that $\forall x \in \mathbb{R}^d$ satisfy $\text{supp}_{W_\Delta^2}(x)$ and $\max_{i \in W_\Delta^2} Q_i(x) \leq \delta''$, we have $\|x - x^*\| \leq \epsilon''$.

Thus for $\Delta \leq \min\{\epsilon_k, \delta''\}$, the first $k$ elements in $W_\Delta$ will be $W^2$, and any coordinate $i \notin W^2$ can not be included in $W_\Delta$. Therefore $W_\Delta = W^2$.

\[\Box\]
G Proof of Theorem 5

Proof. Given the number of iteration \( t \), denote \( Z_t = \{ i \in [d] \mid x^t_i = 0 \ \forall t' < t \} \), which is the entries of \( x^t \) that filled with 0’s. and \( \mathcal{V}_t = \{ i \in [d] \mid |\nabla_i f(x^t) - \nabla_i f(x^*)| \leq \delta_i \ \forall t' \geq t \} \).

From Lemma 3 (in the main text), we know that any coordinates in \( Z_t \cap \mathcal{V}_t \) will always stay at 0 and thus cannot be in \( W \); that is

\[
\mathcal{V}_t \subset [d](Z_t \cap \mathcal{V}_t) \ \forall t > 0
\Rightarrow |W| \leq \min_{t \in [d]} \{ d - |Z_t \cap \mathcal{V}_t| \}.
\] (11)

Recall the definition of the set of good steps until the \( t \)-th iteration \( G_t \subset \[t\].

\[
|\mathcal{V}_t| = \sum_{i=1}^d 1\{ |\nabla_i f(x^{t'}) - \nabla_i f(x^*)| \leq \delta_i \ \forall t' \geq t \}
\geq \sum_{i=1}^d 1\{ \|\nabla f(x^{t'}) - \nabla f(x^*)\|_\infty \leq \delta_i \ \forall t' \geq t \}
\overset{(i)}{=} \sum_{i=1}^d 1\{ L_\infty \|x^t - x^*\|_1 \leq \delta_i \ \forall t' \geq t \}
\geq \sum_{i=1}^d 1\{ L_\infty \sup_{t' \geq t} \|x^{t'} - x^*\|_1 \leq \delta_i \},
\] (12)

where (i) follows from the \( l_\infty \) smoothness assumption.

By the definition of \( G_t \) in section A, we also have \( |Z_t| \geq d - |\mathcal{G}_t| \), and further

\[
|Z_t \cap \mathcal{V}_t| = |Z_t| + |\mathcal{V}_t| - |Z_t \cup \mathcal{V}_t|
\geq d - |\mathcal{G}_t| + |\mathcal{V}_t| - d
\geq |\mathcal{V}_t| - |\mathcal{G}_t|.
\] (13)

Plug the above result in Eq. (11), we get

\[
|W| \leq \min_{t > 0} \{ d - |\mathcal{V}_t| + |\mathcal{G}_t| \}
\leq \min_{t > 0} \left\{ d - \sum_{i=1}^d 1\{ L_\infty \sup_{t' \geq t} \|x^{t'} - x^*\|_1 \leq \delta_i \} + |\mathcal{G}_t| \right\}
\leq \min_{t \in [d]} \left\{ d - \sum_{i=1}^d 1\{ L_\infty \sup_{t' \geq t} \|x^{t'} - x^*\|_1 \leq \delta_i \} + t \right\}
= \min_{t \in [d]} B_t + t,
\] (14)

where \( B_t \) is defined as \( B_t := d - p_\delta \{ L_\infty \sup_{t \geq t} \{ \|x^t - x^*\|_1 \} \} \) in Theorem 3. \( \square \)
H Proof of Corollary 6

Proof. Similar to the proof of Theorem 5, denote \( Z_t = \{ i \in [d] \mid x'_i = 0 \ \forall t' < t \} \), which is the entries of \( x' \) that filled with 0’s. and \( V_t = \{ i \in [d] \mid \| \nabla_i f(x') - \nabla_i f(x^*) \| \leq \delta_i \ \forall t' \geq t \} \).

From Lemma 3 (in the main text), we know that any coordinates in \( Z_t \cap V_t \) will always stay at 0 and thus cannot be in \( W_t \), that is

\[
W \subset [d] \setminus (Z_t \cap V_t) \ \forall t > 0
\Rightarrow |W| \leq \min_{t \in [d]} \{ d - |Z_t \cap V_t| \}.
\]

(15)

Recall the definition of the set of good steps until the \( t \)-th iteration \( G_t \subset [t] \).

\[
|V_t| = \sum_{i=1}^{d} 1\{ |\nabla_i f(x') - \nabla_i f(x^*)| \leq \delta_i \ \forall t' \geq t \}
\]

\[
\geq \sum_{i=1}^{d} 1\{ \| \nabla f(x') - \nabla f(x^*) \|_\infty \leq \delta_i \ \forall t' \geq t \}
\]

(i)

\[
\geq \sum_{i=1}^{d} 1\{ L_\infty \| x' - x^* \|_1 \leq \delta_i \ \forall t' \geq t \}
\]

(ii)

\[
\geq \sum_{i=1}^{d} 1\{ L_\infty \sqrt{2 \mu_1} (F(x') - F(x^*)) \leq \delta_i \ \forall t' \geq t \}
\]

(iii)

\[
\geq \sum_{i=1}^{d} 1\{ L_\infty \sqrt{2 \mu_1} \left( F(x') - F(x^*) \right) \leq \delta_i \}
\]

(iv)

\[
= p_\delta \left( L_\infty \sqrt{2 \mu_1} \left( F(x') - F(x^*) \right) \right)
\]

(v)

\[
\geq p_\delta \left( L_\infty \sqrt{2 \mu_1} \left( F(x') - F(x^*) \right) \right)
\]

(16)

where (i) follows from the \( l_\infty \) smoothness assumption, (ii) is from \( \mu_1 \) strongly convex, (iii) is true since \( F(x') \) is a decreasing sequence, (iv) is by the definition of \( p_\delta(\cdot) \), (v) directly follows from Theorem 4.

By the definition of \( G_t \), we also have \( |Z_t| \geq d - |G_t| \), and further

\[
|Z_t \cap V_t| = |Z_t| + |V_t| - |Z_t \cup V_t|
\]

\[
\geq d - |G_t| + |V_t| - d
\]

\[
\geq |V_t| - |G_t|.
\]

(17)
Plug the above result in Eq. (15), we get

\[ |W| \leq \min_{t > 0} \left\{ d - |\mathcal{V}_t| + |\mathcal{G}_t| \right\} \]

\[ \leq \min_{t > 0} \left\{ d - \left( L_\infty \sqrt{\frac{2 |\mathcal{G}_t|}{\mu_1} \prod_{i=1}^{t-1} \left( 1 - \frac{\mu_1^{(\tau+i-1)}}{L} \right) (F(0) - F^*)} + |\mathcal{G}_t| \right) \right\} \]

\[ \leq \min_{t \in [d]} \left\{ d - \left( L_\infty \sqrt{\frac{2 \prod_{i=1}^{t} \left( 1 - \frac{\mu_1^{(\tau+i-1)}}{L} \right) (F(0) - F^*)} + t \right) \right\} = \min_{t \in [d]} B_t + t, \]

(18)

where \( B_t \) is defined as \( B_t := d - p_\delta \left( \sqrt{\frac{2L_\infty^2}{\mu_1} \prod_{i=0}^{t-1} \left( 1 - \frac{\mu_1^{(\tau+i-1)}}{L} \right) (F(0) - F^*)} \right) \) in Theorem 5.

References

