Greed Meets Sparsity: Understanding and Improving Greedy Coordinate Descent for Sparse Optimization

Huang Fang, Zhenan Fan, Yifan Sun, Michael P. Friedlander

Appendix

A Prelimiaries

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. (2) 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 $\mathcal{G}_t$, since a bad step always follows a good step, it is easy to verify that

$$|\mathcal{G}_t| \leq \left\lceil \frac{t}{2} \right\rceil. \tag{1}$$

Recall the selection rule in section 3:
Selection rule 1 (GS-s rule)**.** *Select* $i \in \arg\max_j Q_j(x^t)$ *where*

$$Q_i(x) = \min_{s \in \partial g_i} |\nabla_i f(x) + s|. \tag{2}$$

Lemma 2 ([Karimireddy et al., 2019])**.** *Assume* $f(\cdot)$ *is* μ_1 *strongly convex with respect to 1-norm, then the iterates generated from Algorithm 1 with GS-s rule (selection rule 1) satisfy*

$$F(x^t) - F(x^*) \leq \left(1 - \frac{\mu_1}{L}\right)^{\lceil \frac{t}{2} \rceil} (F(0) - F(x^*)).$$

The above lemma is from [Karimireddy et al., 2019].
Lemma 3 ([Karimireddy et al., 2019])**.** *Consider* $g(\cdot)$ *to be* ℓ_1 *regularization or non-negative constraint. Then if the* t*-th iteration is a good step, we have*

$$F(x^{t+1}) \leq F(x^t) - \frac{1}{2L} \max_{i \in [d]} Q_i(x^t)^2, \tag{3}$$

where $Q_i(\cdot)$ *is defined in the GS-s rule (selection rule 1).*

B Proof of Theorem Sketch 2

Proof. Let $W = \{w_1, w_2, ..., w_k\}$ s.t. $w_1 < w_2 < ... < w_k \in \mathbf{N}$, 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 $i \in [k], y \in \mathbb{R}^k$,

$$\begin{aligned} h(y+\alpha e_i) &= f(\sum_{j=1}^{k} y_j e_{w_j} + \alpha e_{w_i}) \\ &\leq f(\sum_{j=1}^{k} y_j e_{w_j}) + \alpha \nabla_{w_i} f(\sum_{j=1}^{k} y_j e_{w_j}) + \frac{L}{2}\alpha^2 \\ &= h(y) + \alpha \nabla_i h(y) + \frac{L}{2}\alpha^2 \end{aligned} \tag{4}$$

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 $x^t = \sum_{i=1}^{k} e_{w_i} y_i^t \ \forall t \geq 0$. We prove by induction:

When $t = 0$, obviously we have $x^0 = \sum_{i=1}^{k} e_{w_i} y_i^0 = 0$.

Suppose that $x^t = \sum_{i=1}^{k} e_{w_i} y_i^t$, $i = \arg\max_j Q_j(x^t)$, $\tilde{i} = \arg\max_j Q_j(y^t)$ and $i = w_m$, we can show that $\tilde{i} = m$:

Note that

$$\begin{aligned} Q_i(x^t) &= \min_{s \in \partial g_i} |\nabla_i f(x^t) + s| \\ &= \min_{s \in \partial g_m} |\nabla_m h(x) + s| \\ &= Q_m(y^t), \end{aligned} \tag{5}$$

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

$$x_i^{t+\frac{1}{2}} = x_i^t - \frac{1}{L}\nabla f_i(x^t) = y_m^t - \frac{1}{L}\nabla h_m(y^t) = y_m^{t+\frac{1}{2}}.$$

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

$$x_i^{t+1} = \text{prox}_{\frac{1}{L}g_i}\left[x_i^{t+\frac{1}{2}}\right] = \text{prox}_{\frac{1}{L}g_{w_m}}\left[y_m^{t+\frac{1}{2}}\right] = y_{m+1}^{t+1}.$$

Thus we have $x^t = \sum_{i=1}^{k} e_{w_i} y_i^t \ \forall t = 0, 1, 2, ...$

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 μ_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.

□

C Proof of Lemma 3

Proof. If i is not select by Algorithm 1 at the t-th iteration, then $x_i^{t+1} = 0$ 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

$$\begin{aligned} &-\delta_i + \nabla_i f(x^*) \leq \nabla_i f(x^t) \leq \delta_i + \nabla_i f(x^*) \\ \overset{\text{(i)}}{\Rightarrow} &-u_i \leq \nabla_i f(x^t) \leq -l_i, \end{aligned} \tag{6}$$

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

$$\text{prox}_{\frac{g}{L_i}}\left(0 - \frac{1}{L_i}\nabla_i f(x^t)\right) = \arg\min_y \left\{\frac{1}{2}\left(y - \left(-\frac{1}{L_i}\nabla f_i(x^t)\right)\right)^2 + \frac{1}{L_i}g_i(y)\right\} \tag{7}$$

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

$$0 \in y + \frac{1}{L_i}\nabla_i f(x^t) + \frac{1}{L_i}\partial g_i(y) \tag{8}$$

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

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

□

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: $|\mathcal{G}_t| = \{i_1, i_2, \ldots, i_k\}$, where $k \geq \lceil \frac{t}{2} \rceil$.

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 μ_1 strongly convex w.r.t. 1-norm and $\tau + m - 1$–sparse vectors. Thus $\forall y \in \mathbb{R}^d$ that is τ-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

$$F(y) \geq F(x^{i_m}) + \langle \partial F(x^{i_m}), y - x^{i_m}\rangle + \frac{\mu_1^{(\tau+m-1)}}{2}\|y - x^{i_m}\|_1^2, \tag{9}$$

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 τ sparse,

$$\begin{aligned} F(x^*) &\geq F(x^{i_m}) - \sup_{\|y\|_0 \leq \tau}\left(\langle -\partial F(x^{i_m}), y - x^{i_m}\rangle - \frac{\mu_1^{(\tau+m-1)}}{2}\|y - x^{i_m}\|_1^2\right) \\ &\geq F(x^{i_m}) - \sup_{y \in \mathbb{R}^d}\left(\langle -\partial F(x^{i_m}), y - x^{i_m}\rangle - \frac{\mu_1^{(\tau+m-1)}}{2}\|y - x^{i_m}\|_1^2\right) \\ &\overset{\text{(i)}}{=} F(x^{i_m}) - \left(\frac{\mu_1^{(\tau+m-1)}}{2}\|\cdot\|_1^2\right)^*(-\partial F(x^{i_m})) \\ &\overset{\text{(ii)}}{=} F(x^{i_m}) - \frac{1}{2\mu_1^{(\tau+m-1)}}\|\partial F(x^{i_m})\|_\infty^2, \end{aligned}$$

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

More specifically,

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

By the definition of $Q_i(\cdot)$ in the GS-s rule (selection rule 1), we further have

$$F(x^*) \geq F(x^{i_m}) - \frac{1}{2\mu_1^{(\tau+m-1)}} \max_{i\in[d]} Q_i(x^{i_m})^2. \tag{10}$$

Recall Lemma 3, we have

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

Plug the above equation into Eq. (10)

$$\begin{aligned} F(x^*) &\geq F(x^{i_m}) - \frac{L}{\mu_1^{(\tau+m-1)}}(F(x^{i_m+1}) - F(x^{i_m})) \\ \Rightarrow F(x^{i_m+1}) - F^* &\leq \left(1 - \frac{\mu_1^{(\tau+m)}}{L}\right)(F(x^{i_m}) - F^*). \end{aligned}$$

By applying the above inequality recursively, we get

$$\begin{aligned} F(x^t) - F^* &\leq \prod_{m=1}^{k}\left(1 - \frac{\mu_1^{(\tau+m-1)}}{L}\right)(F(0) - F^*) \\ &\leq \prod_{i=1}^{\lceil\frac{t}{2}\rceil}\left(1 - \frac{\mu_1^{(\tau+i-1)}}{L}\right)(F(0) - F^*), \end{aligned}$$

which completes the proof.

□

E Proof of Theorem 9

Proof. This proof is essentially the same as Theorem 4, the difference is that, by the definition of the Δ-GS-s rule (selection rule 8), 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 $|\text{supp}(x^*) \cup \text{supp}(x^t)| \leq |W_\Delta|\ \forall t > 0$. Then we can incorporate the new Lemma into the analysis of Theorem 4 and get

$$\begin{aligned} F(x^t) - F^* &\leq \left(1 - \frac{\Delta\mu_1^{(|W_\Delta|)}}{L}\right)^{\lceil\frac{t}{2}\rceil}(F(0) - F^*) \\ &\leq \left(1 - \frac{\Delta\mu_2}{|W_\Delta|L}\right)^{\lceil\frac{t}{2}\rceil}(F(0) - F^*)\,. \end{aligned}$$

□

F Proof of Theorem 10

Proof.

Clarify some notations

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

We denote the t-th iterate from the Δ-GCD algorithm as x^t and the t-th iterate from the totally corrective greedy algorithm (TCGA) as $\tilde{x}^t$. $W^\sharp = \{\tilde{i}_1, \tilde{i}_2, \ldots, \tilde{i}_k\}$, 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_Δ is the same as the first j elements in $W^\sharp$.

We prove this claim by induction, when $j = 1$, $\forall \Delta \leq 1$, Δ-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 $\tilde{x}^j$:

$$\tilde{x}^j := \operatorname*{arg\,min}_{\mathrm{supp}(x)\subseteq W_j} f(x) + g(x)$$

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

Denote $Q_i(x^t)$ (recall x^t is generated from Δ-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_Δ at some iteration t, we have

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

also by the induction assumption, we know that $\mathrm{supp}(x^t) \subseteq W_j$. Putting these two conditions together, we get $\|x^t - x^j\| \leq \epsilon'$ 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_Δ is just $W^\sharp$.

By the nondegeneracy assumption i.e., $\delta_i > 0$ $\forall x_i^* = 0$ 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 $Q_i(x) = 0$ $\forall i \notin W^\sharp$ (note that $\mathrm{supp}(x^*) \in W^\sharp$).

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

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

□

G Proof of Theorem 6

Proof. Given the number of iteration t, denote $\mathcal{Z}_t = \{i \in [d] \mid x_i^{t'} = 0 \ \forall t' < t\}$, 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 $\mathcal{Z}_t \cap \mathcal{V}_t$ will always stay at 0 and thus cannot be in W, that is

$$\begin{aligned} & W \subset [d] \backslash (\mathcal{Z}_t \cap \mathcal{V}_t) \ \ \forall t > 0 \\ \Rightarrow & |W| \leq \min_{t \in [d]} \{d - |\mathcal{Z}_t \cap \mathcal{V}_t|\} \end{aligned} \tag{11}$$

Recall the definition of the set of good steps until the t-th iteration $\mathcal{G}_t \subset [t]$.

$$\begin{aligned} |\mathcal{V}_t| &= \sum_{i=1}^{d} \mathbf{1}\{|\nabla_i f(x^{t'}) - \nabla_i f(x^*)| \leq \delta_i \ \ \forall t' \geq t\} \\ &\geq \sum_{i=1}^{d} \mathbf{1}\{\|\nabla f(x^{t'}) - \nabla f(x^*)\|_\infty \leq \delta_i \ \ \forall t' \geq t\} \\ &\overset{\text{(i)}}{\geq} \sum_{i=1}^{d} \mathbf{1}\{L_\infty \|x^{t'} - x^*\|_1 \leq \delta_i \ \ \forall t' \geq t\} \\ &\geq \sum_{i=1}^{d} \mathbf{1}\{L_\infty \sup_{t' \geq t} \|x^t - x^*\|_1 \leq \delta_i\} \end{aligned} \tag{12}$$

where (i) follows from the l_∞ smoothness assumption.

By the definition of $\mathcal{G}_t$ in section A, we also have $|\mathcal{Z}_t| \geq d - |\mathcal{G}_t|$, and further

$$\begin{aligned} |\mathcal{Z}_t \cap \mathcal{V}_t| &= |\mathcal{Z}_t| + |\mathcal{V}_t| - |\mathcal{Z}_t \cup \mathcal{V}_t| \\ &\geq d - |\mathcal{G}_t| + |\mathcal{V}_t| - d \\ &\geq |\mathcal{V}_t| - |\mathcal{G}_t| \end{aligned} \tag{13}$$

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

$$\begin{aligned} |W| &\leq \min_{t>0} \{d - |\mathcal{V}_t| + |\mathcal{G}_t|\} \\ &\leq \min_{t>0} \left\{ d - \sum_{i=1}^{d} \mathbf{1}\{L_\infty \sup_{t' \geq t} \|x^{t'} - x^*\|_1 \leq \delta_i\} + |\mathcal{G}_t| \right\} \\ &\leq \min_{t \in [d]} \left\{ d - \sum_{i=1}^{d} \mathbf{1}\{L_\infty \sup_{t' \geq t} \|x^{t'} - x^*\|_1 \leq \delta_i\} + t \right\} \\ &= \min_{t \in [d]} B_t + t \end{aligned} \tag{14}$$

where B_t is defined as $B_t := d - p_\delta \left(L_\infty \sup_{i \geq t} \left\{ \|x^i - x^*\|_1 \right\} \right)$ in Theorem 6 □

H Proof of Corollary 7

Proof. Similar to the proof of Theorem 6, denote $\mathcal{Z}_t = \{i \in [d] \mid x_i^{t'} = 0 \ \forall t' < t\}$, 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 $\mathcal{Z}_t \cap \mathcal{V}_t$ will always stay at 0 and thus cannot be in W, that is

$$\begin{aligned} & W \subset [d] \backslash (\mathcal{Z}_t \cap \mathcal{V}_t) \quad \forall t > 0 \\ \Rightarrow & |W| \leq \min_{t \in [d]} \{d - |\mathcal{Z}_t \cap \mathcal{V}_t|\} \end{aligned} \tag{15}$$

Recall the definition of the set of good steps until the t-th iteration $\mathcal{G}_t \subset [t]$.

$$\begin{aligned} |\mathcal{V}_t| &= \sum_{i=1}^{d} \mathbf{1}\{|\nabla_i f(x^{t'}) - \nabla_i f(x^*)| \leq \delta_i \quad \forall t' \geq t\} \\ &\geq \sum_{i=1}^{d} \mathbf{1}\{\|\nabla f(x^{t'}) - \nabla f(x^*)\|_\infty \leq \delta_i \quad \forall t' \geq t\} \\ &\overset{\text{(i)}}{\geq} \sum_{i=1}^{d} \mathbf{1}\{L_\infty \|x^{t'} - x^*\|_1 \leq \delta_i \quad \forall t' \geq t\} \\ &\overset{\text{(ii)}}{\geq} \sum_{i=1}^{d} \mathbf{1}\left\{L_\infty \sqrt{\frac{2}{\mu_1}(F(x^t) - F(x^*))} \leq \delta_i \quad \forall t' \geq t\right\} \\ &\overset{\text{(iii)}}{=} \sum_{i=1}^{d} \mathbf{1}\left\{L_\infty \sqrt{\frac{2}{\mu_1}(F(x^t) - F(x^*))} \leq \delta_i\right\} \\ &\overset{\text{(iv)}}{=} p_\delta\left(L_\infty \sqrt{\frac{2}{\mu_1}(F(x^t) - F(x^*))}\right) \\ &\overset{\text{(v)}}{\geq} p_\delta\left(L_\infty \sqrt{\frac{2}{\mu_1} \prod_{i=1}^{|\mathcal{G}_t|} \left(1 - \frac{\mu_1^{(\tau+i-1)}}{L}\right)(F(0) - F^*)}\right) \end{aligned} \tag{16}$$

where (i) follows from the l_∞ smoothness assumption, (ii) is from μ_1 strongly convex, (iii) is true since $F(x^t)$ is a decreasing sequence, (iv) is by the definition of $p_\delta(\cdot)$, (v) directly follows from Theorem 4.

By the definition of $\mathcal{G}_t$, we also have $|\mathcal{Z}_t| \geq d - |\mathcal{G}_t|$, and further

$$\begin{aligned} |\mathcal{Z}_t \cap \mathcal{V}_t| &= |\mathcal{Z}_t| + |\mathcal{V}_t| - |\mathcal{Z}_t \cup \mathcal{V}_t| \\ &\geq d - |\mathcal{G}_t| + |\mathcal{V}_t| - d \\ &\geq |\mathcal{V}_t| - |\mathcal{G}_t|. \end{aligned} \tag{17}$$

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

$$
\begin{aligned}
|W| &\leq \min_{t>0} \{d - |\mathcal{V}_t| + |\mathcal{G}_t|\} \\
&\leq \min_{t>0} \left\{ d - \left(L_\infty \sqrt{\frac{2}{\mu_1} \prod_{i=1}^{|\mathcal{G}_t|} \left(1 - \frac{\mu_1^{(\tau+i-1)}}{L}\right) (F(0) - F^*)} \right) + |\mathcal{G}_t| \right\} \\
&\leq \min_{t \in [d]} \left\{ d - \left(L_\infty \sqrt{\frac{2}{\mu_1} \prod_{i=1}^{t} \left(1 - \frac{\mu_1^{(\tau+i-1)}}{L}\right) (F(0) - F^*)} \right) + t \right\} \\
&= \min_{t \in [d]} B_t + t
\end{aligned} \tag{18}
$$

where B_t is defined as $B_t := d - p_\delta \left(\sqrt{\frac{2L_\infty^2}{\mu_1} \prod_{i=0}^{t-1} \left(1 - \frac{\mu_1^{(\tau+i)}}{L}\right) (F(0) - F^*)} \right)$ in Theorem 6 □

References

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

[Karimireddy et al., 2019] Karimireddy, S. P., Koloskova, A., Stich, S. U., and Jaggi, M. (2019). Efficient greedy coordinate descent for composite problems. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2887–2896. PMLR.

[Nutini et al., 2015] Nutini, J., Schmidt, M. W., Laradji, I. H., Friedlander, M. P., and Koepke, H. A. (2015). Coordinate descent converges faster with the gauss-southwell rule than random selection. In *Proceedings of ICML*, pages 1632–1641.