Supplementary Materials

Huang Fang, Guanhua Fang, Tan Yu, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, Washington 98004, USA
{fangazq877, fanggh2018, Tanyuuynat, pingli98}@gmail.com

1 PROOFS

Proof of Lemma 6.2

Proof. Given $\mathbf{x} \in \mathbb{R}^d$,

$$\begin{aligned}
\|\mathbf{x}\|_{\mathcal{B},1} &:= \sup_{\|\mathbf{z}\|_{\mathcal{B},\infty} \le 1} \langle \mathbf{z}, \mathbf{x} \rangle \\
&= \sup_{\mathbf{z} \in \mathbb{R}^d} \left\{ \langle \mathbf{z}, \mathbf{x} \rangle \,\middle|\, \max_{i \in [k]} \frac{1}{\sqrt{|B_i|}} \|\mathbf{z}_{B_i}\|_2 \le 1 \right\} \\
&= \sup_{\mathbf{z} \in \mathbb{R}^d} \left\{ \sum_{i=1}^k \langle \mathbf{z}_{B_i}, \mathbf{x}_{B_i} \rangle \,\middle|\, \frac{1}{\sqrt{|B_i|}} \|\mathbf{z}_{B_i}\|_2 \le 1 \; \forall i \in [k] \right\} \\
&= \sum_{i=1}^k \sup_{\mathbf{z} \in \mathbb{R}^d} \left\{ \langle \mathbf{z}_{B_i}, \mathbf{x}_{B_i} \rangle \,\middle|\, \frac{1}{\sqrt{|B_i|}} \|\mathbf{z}_{B_i}\|_2 \le 1 \right\} \\
&\overset{\text{(i)}}{=} \sum_{i=1}^k \sqrt{|B_i|} \|\mathbf{x}_{B_i}\|_2.
\end{aligned}$$

For (i), the maximum is attained when $\mathbf{z}_{B_i} = \sqrt{|B_i|}\mathbf{x}_{B_i} / \|\mathbf{x}_{B_i}\|_2$. □

Proof of Theorem 6.5

Proof. We begin with Equation (6.1),

$$\text{Equation (6.1)} \le f(\mathbf{x}^{(t)}) - \frac{1}{2L_{\max}} \left(\frac{\|\nabla f(\mathbf{x}^{(t)})\|_{\mathcal{B},\infty}}{\|\nabla f(\mathbf{x}^{(t)})\|_\infty} \right)^2 \|\nabla f(\mathbf{x}^{(t)})\|_\infty^2$$

The next step follows from the refined analysis of GCD from Nutini et al. (2015), we present it here for completeness. Since μ_1 is strongly convex, we have

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu_1}{2} \|\mathbf{x} - \mathbf{y}\|_1^2, \; \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

By minimizing left-hand and right-hand sides over $\mathbf{x}$, we get

$$
\begin{aligned}
f^* &\geq f(y) - \sup_{\mathbf{x}\in\mathbb{R}^d}\left(\langle \nabla f(\mathbf{y}), \mathbf{y}-\mathbf{x}\rangle - \frac{\mu_1}{2}\|\mathbf{y}-\mathbf{x}\|_1^2\right) \\
&= f(y) - \left(\frac{\mu_1}{2}\|\cdot\|_1^2\right)^*(\nabla f(\mathbf{y})) \\
&\overset{\text{(i)}}{=} f(y) - \frac{1}{2\mu_1}\|\nabla f(y)\|_\infty^2, \qquad (1.1)
\end{aligned}
$$

where (i) uses the fact that the convex conjugate of $\frac{1}{2}\|\cdot\|_1^2$ is $\frac{1}{2}\|\cdot\|_\infty^2$. By subtracting f^* from left-hand and right-hand sides of Eq. (6.1) and combining with Eq. (1.1), we get

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}^{(t)}) - f^*\right] \\
&\leq \left(1 - \frac{\mu_1}{L_{\max}}\frac{\|\nabla f(\mathbf{x}^{(t)})\|_{\mathcal{B},\infty}^2}{\|\nabla f(\mathbf{x}^{(t)})\|_\infty^2}\right)\left(f(\mathbf{x}^{(t)}) - f^*\right). \qquad (1.2)
\end{aligned}
$$

Furthermore, with $\frac{\|x\|_2^2}{d} \leq \|x\|_{\mathcal{B},\infty}^2$ and Eq. (6.1), we get

$$
\mathbb{E}\left[f(\mathbf{x}^{(t)})\right] \leq f(\mathbf{x}^{(t)}) - \frac{1}{2dL_{\max}}\|\nabla f(\mathbf{x}^{(t)})\|_2^2, \qquad (1.3)
$$

Using the same argument to derive Eq. (1.2) or following the standard analysis for randomized coordinate descent, we get

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}^{(t)}) - f^*\right] \\
&\leq \left(1 - \frac{\mu_2}{L_{\max}}\frac{\|\nabla f(\mathbf{x}^{(t)})\|_{\mathcal{B},\infty}^2}{\|\nabla f(\mathbf{x}^{(t)})\|_2^2}\right)\left(f(\mathbf{x}^{(t)}) - f^*\right). \qquad (1.4)
\end{aligned}
$$

We complete the proof by combining Eq. (1.2) and Eq. (1.4). □

Proof of Theorem 6.6

Proof. We begin with Equation (6.1) and follow the standard proof template (Karimireddy et al., 2019; Dhillon et al., 2011),

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{(t+1)}) \mid \mathbf{x}^{(t)}] &\leq f(\mathbf{x}^{(t)}) - \frac{\eta^2}{2L_{\max}}\|\nabla f(\mathbf{x}^{(t)})\|_\infty^2 \\
&\overset{\leq}{\text{(i)}} f(\mathbf{x}^{(t)}) - \frac{\eta^2}{2L_{\max}\|\mathbf{x}^{(t)} - \mathbf{x}^*\|_1^2}(f(\mathbf{x}^{(t)}) - f^*)^2. \\
&\leq f(\mathbf{x}^{(t)}) - \frac{\eta^2}{2L_{\max}D^2}(f(\mathbf{x}^{(t)}) - f^*)^2.
\end{aligned}
$$

where (i) is from the following inequality

$$
f(\mathbf{x}^{(t)}) - f^* \leq \langle \mathbf{x}^* - \mathbf{x}^{(t)}, -\nabla f(\mathbf{x}^{(t)})\rangle \leq \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_1\|\nabla f(\mathbf{x}^{(t)})\|_\infty.
$$

Taking expectation on both sides,

$$
\mathbb{E}[f(\mathbf{x}^{(t+1)})] \leq \mathbb{E}[f(\mathbf{x}^{(t)})] - \frac{\eta^2}{2L_{\max}D^2}(\mathbb{E}[f(\mathbf{x}^{(t)})] - f^*)^2,
$$

Note that we use the fact that $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$ to derive the above property. Denote $\mathbb{E}[f(\mathbf{x}^{(t)})] - f^*$ as h_t, then we can get

$$
h_{t+1} \leq h_t - \frac{\eta^2}{2L_{\max}D^2}h_t^2. \qquad (1.5)
$$

Dividing both side by $h_{t+1}h_t$, we get

$$\frac{1}{h_t} \le \frac{1}{h_{t+1}} - \frac{\eta^2}{2L_{\max}D^2}\frac{h_t}{h_{t+1}} \overset{\text{(i)}}{\le} \frac{1}{h_{t+1}} - \frac{\eta^2}{2L_{\max}D^2}, \tag{1.6}$$

where (i) is from the fact that $\{h_t\}_{t=1}^{\infty}$ is a decreasing sequence and $h_t/h_{t+1} \ge 1$. Summing Equation (1.6) over $t \in \{0, 1, \dots, T\}$, we get

$$\begin{aligned} \frac{1}{h_0} - \frac{1}{h_T} &\le -\frac{T\eta^2}{2L_{\max}D^2} \\ \Longrightarrow h_T &\le \frac{2L_{\max}D^2}{\eta^2 T}, \end{aligned}$$

which completes the proof.

□

Proof of Theorem 6.7

Proof. Given any vector r, we let $z_i = \mathbf{a}_i^T r$ and define $m_j := \boldsymbol{\mu}_j^T r$. Therefore,

$$\sum_{i \in B_j} z_i^2 = (m_j + z_i - m_j)^2 = |B_j| m_j^2 + 2m_j \sum_{i \in B_j}(z_i - m_j) + \sum_{i \in B_j}(z_i - m_j)^2. \tag{1.7}$$

According to Rudelson and Vershynin (2010), with probability at least $1 - 2\exp\{-n/2\}$, we have that

$$\sum_{i \in B_j}(z_i - m_j)^2 = \|\tilde{A}_{B_j} r\|^2 \ge (\sqrt{|B_j|} - 2\sqrt{n})^2 \|r\|_2^2 \sigma^2$$

and

$$|\sum_{i \in B_j}(z_i - m_j)| \le \sigma \|r\|_2 \sqrt{|B_j| n \log n}$$

hold for all r. Here $\tilde{A}_{B_j}$ is the jth-block submatrix of A by shifting mean to zero. Therefore, we have

$$\sum_{i \in B_j} z_i^2 \ge |B_j| m_j^2 - 2m_j \sigma \|r\|_2 \sqrt{|B_j| n \log n} + (\sqrt{|B_j|} - 2\sqrt{n})^2 \|r\|_2^2 \sigma^2 \tag{1.8}$$

by simplifying (1.7). On the other hand,

$$\max_{i \in B_j} |z_i| = \max_{i \in B_j} |\boldsymbol{\mu}_j^r + (\mathbf{a}_i - \boldsymbol{\mu}_j)^T r| \le m_j + \max_{i \in B_j} |(\mathbf{a}_i - \boldsymbol{\mu}_j)^T r| \le m_j + 2\log|B_j|\sqrt{n}\|r\|_2 \sigma$$

holds with probability at least $1 - 2|B_j|\exp\{-n/4\}$. Thus, when $|B_j| \gg n$, we get

$$\begin{aligned} & \frac{\sqrt{\sum_{i \in B_j} z_i^2}}{\max_{i \in B_j} |z_j|} \\ \ge\ & \frac{\sqrt{|B_j| m_j^2 - 2m_j \sigma \|r\|_2 \sqrt{|B_j| n \log n} + (\sqrt{|B_j|} - 2\sqrt{n})^2 \|r\|_2^2 \sigma^2}}{m_j + 2\log|B_j|\sqrt{n}\|r\|_2 \sigma} \\ \ge\ & \frac{c\sqrt{|B_j|}\sqrt{m_j^2 + \|r\|_2^2 \sigma^2}}{\log|B_j|\sqrt{n}\sqrt{m_j^2 + \|r\|_2^2 \sigma^2}} \qquad (1.9) \\ =\ & c\sqrt{|B_j|}/(\log|B_j|\sqrt{n}) \qquad (1.10) \end{aligned}$$

for some universal constant c. Notice that the above results hold for all r. This implies that

$$\frac{\|\nabla_{B_j} f\|_2 / \sqrt{|B_j|}}{\|\nabla_{B_j} f\|_\infty} \ge \frac{c}{\log|B_j|\sqrt{n}}, \tag{1.11}$$

if we specifically take $r = f'(A\mathbf{x}^t)$. This further gives

$$\|\nabla f(\mathbf{x}^t)\|_{\mathcal{B},\infty} \geq \frac{c}{\max_j \log|B_j|\sqrt{n}}\|\nabla f(\mathbf{x}^t)\|_\infty,$$

When $\max_j |B_j| \geq d/k \gg n$, there is huge improvement in lower bound, from $1/(\max_j \sqrt{|B_j|})$ to $1/(\max_j \log|B_j|\sqrt{n})$. This concludes the proof. □

Proof of Theorem 6.10

Proof. We first show that $\|\mathbf{c}_j - \boldsymbol{\mu}_j\| \leq \delta\sqrt{n}$. We compare the difference between kth coordinates of $\mathbf{c}_j$ and $\boldsymbol{\mu}_j$. Then

$$\begin{aligned}
|\mathbf{c}_j(k) - \boldsymbol{\mu}_j(k)| &= \frac{1}{|B_j|}|\sum_{i\in B_j} A_{ik} - \boldsymbol{\mu}_j(k)| \\
&\leq \frac{1}{|B_j|}(|\sum_{i\in B_j\cap B_j^*}(A_{ik} - \boldsymbol{\mu}_j(k))| + |\sum_{i\in B_j\cap B_j^{*c}}(A_{ik} - \boldsymbol{\mu}_j(k))|) \\
&\leq \frac{C_1\log|B_j|\sqrt{|B_j\cap B_j^*|}}{|B_j|} + \frac{|B_j\cap B_j^{*c}|}{|B_j|}(\mu_{gap} + \sigma\log|B_j|), \qquad (1.12) \\
&:= \delta_j.
\end{aligned}$$

holds with probability at least $1 - \frac{1}{|B_j|}$, where $\mu_{gap} := \max_{k\in[n]}\max_{j_1\neq j_2}|\boldsymbol{\mu}_{j_1}(k) - \boldsymbol{\mu}_{j_2}(k)|$ and B_j^{*c} is the complement of B_j^*. By assumption *A1*, it can be checked that $\delta_j \leq 2\sigma$ when $|B_j| \gg n$. Here, (1.12) holds since that $\sum_{i\in B_j\cap B_j^{*c}}(A_{ik} - \boldsymbol{\mu}_j(k))$ is a Gaussian random variable which is $O_p(\sqrt{|B_j\cap B_j^*|})$. For each $i \in B_j\cap B_j^{*c}$, the difference between A_{ik} and $\boldsymbol{\mu}_j(k)$ is at most $|\boldsymbol{\mu}_{j_i}(k) - \boldsymbol{\mu}_j(k)|$ plus noise term, which is further bounded by $\mu_{gap} + \sigma\log|B_j|$.

We next compute the lower bound of $\sum_{i\in B_j} z_i^2$. By use of (1.8), we get

$$\sum_{i\in B_j} z_i^2 \geq \sum_{i\in B_j\cap B_j^*} z_i^2 = |B_j\cap B_j^*|m_j^2 - 2m_j\sigma\|r\|_2\sqrt{|B_j\cap B_j^*|n\log n} + (\sqrt{|B_j\cap B_j^*|} - 2\sqrt{n})^2\|r\|_2^2\sigma^2. \qquad (1.13)$$

We further calculate the upper bound of $\max_{i\in B_j}|z_i|$.

$$\begin{aligned}
\max_{i\in B_j}|z_i| &= \max\{\max_{i\in B_j\cap B_j^*}|z_i|, \max_{i\in B_j\cap B_j^{*c}}|z_i|\} \\
&\leq \max\{m_j + 2\log|B_j|\sqrt{n}\|r\|_2\sigma, \max_{i\in B_j\cap B_j^{*c}}|z_i|\} \\
&= \max\{m_j + 2\log|B_j|\sqrt{n}\|r\|_2\sigma, \max_{i\in B_j\cap B_j^{*c}}|\mathbf{c}_j^T r - \boldsymbol{\mu}_j^T r + \boldsymbol{\mu}_j^T r + (\mathbf{a}_i - \mathbf{c}_j)^T r|\} \\
&\leq \max\{m_j + 2\log|B_j|\sqrt{n}\|r\|_2\sigma, \sqrt{n}\delta_j\|r\| + m_j + \max_{i\in B_j\cap B_j^{*c}}|(\mathbf{a}_i - \mathbf{c}_j)^T r|\} \\
&\leq \max\{m_j + 2\log|B_j|\sqrt{n}\|r\|_2\sigma, \sqrt{n}\delta_j\|r\| + m_j + \max_{i\in B_j^*}C|(\mathbf{a}_i - \mathbf{c}_j)^T r|\} \\
&\leq \max\{m_j + 2\log|B_j|\sqrt{n}\|r\|_2\sigma, \sqrt{n}(C+1)\delta_j\|r\| + m_j + 2C\log|B_j^*|\sqrt{n}\|r\|_2\sigma\} \\
&\leq m_j + \sqrt{n}(C+1)\delta_j\|r\| + 2C\log|B_j^*|\sqrt{n}\|r\|_2\sigma. \qquad (1.14)
\end{aligned}$$

By (1.13) and (1.14), when $|B_j| \gg n$, we get

$$\begin{aligned}
& \frac{\sqrt{\sum_{i\in B_j} z_i^2}}{\max_{i\in B_j} |z_j|} \\
\geq \quad & \frac{\sqrt{|B_j \cap B_j^*| m_j^2 - 2m_j\sigma\|r\|_2\sqrt{|B_j\cap B_j^*| n\log n} + (\sqrt{|B_j\cap B_j^*|} - 2\sqrt{n})^2\|r\|_2^2\sigma^2}}{m_j + \sqrt{n}(C+1)\delta_j\|r\| + 2C\log|B_j^*|\sqrt{n}\|r\|_2\sigma} \\
\geq \quad & \frac{c\sqrt{|B_j\cap B_j^*|}\sqrt{m_j^2 + \|r\|_2^2\sigma^2}}{\max\{C\log|B_j^*|, C+1\}\sqrt{n}\sqrt{m_j^2 + \|r\|_2^2\sigma^2 + \|r\|_2^2\delta_j^2}} \\
\geq \quad & c\sqrt{|B_j|}/(C(\log|B_j^*|+1)\sqrt{n}) \qquad (1.15)
\end{aligned}$$

by adjusting some universal constant c. Notice that the above results hold for all r. This implies that

$$\frac{\|\nabla_{B_j} f\|_2/\sqrt{|B_j|}}{\|\nabla_{B_j} f\|_\infty} \geq \frac{c}{C(\log|B_j^*|+1)\sqrt{n}}. \qquad (1.16)$$

This further gives

$$\|\nabla f(\mathbf{x}^t)\|_{\mathcal{B},\infty} \geq \frac{c}{C\max_j(\log|B_j^*|+1)\sqrt{n}}\|\nabla f(\mathbf{x}^t)\|_\infty,$$

When $\max_j |B_j| \geq d/k \gg n$, there is huge improvement in lower bound, from $1/(\max_j \sqrt{|B_j|})$ to $1/(C(\log|B_j^*| + 1)\sqrt{n})$. □

Proof of Theorem 6.11

Our proof follows the same pattern as the proof of ASCD (Lu et al., 2018).

Lemma 1.1. *Define* $\mathbf{s}^{(t+1)} := \mathbf{y}^{(t)} - \frac{1}{dL_{\max}}\nabla f(\mathbf{y}^{(t)})$*, then*

$$\mathbb{E}_t f(\mathbf{x}^{(t)}) \leq f(\mathbf{y}^{(t)}) + \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{s}^{(t+1)} - \mathbf{y}^{(t)}\rangle + \frac{dL_{\max}}{2}\|\mathbf{s}^{(t+1)} - \mathbf{y}^{(t)}\|^2.$$

Proof.

$$\begin{aligned}
\mathbb{E}_t f(\mathbf{x}^{(t+1)}) &\leq f(\mathbf{y}^{(t)}) - \frac{1}{2L_{\max}}\mathbb{E}_t(\nabla_{j_1} f(\mathbf{y}^{(t)}))^2 && (1.17)\\
&\leq f(\mathbf{y}^{(t)}) - \frac{1}{2dL_{\max}}\|\nabla f(\mathbf{y}^{(t)})\|^2 && (1.18)\\
&= f(\mathbf{y}^{(t)}) + \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{s}^{(t+1)} - \mathbf{y}^{(t)}\rangle + \frac{dL_{\max}}{2}\|\mathbf{s}^{(t+1)} - \mathbf{y}^{(t)}\|^2 && (1.19)
\end{aligned}$$

Here, (1.18) holds due to the following fact,

$$\begin{aligned}
\mathbb{E}_t|\nabla_{j_1} f(\mathbf{y}^{(t)})|^2 &= \mathbb{E}_t \max_j |\nabla_{i_j} f(\mathbf{y}^{(t)})|^2 \\
&\geq \mathbb{E}_t \sum_j \frac{|B_j|}{d}|\nabla_{i_j} f(\mathbf{y}^{(t)})|^2 \\
&= \sum_j \frac{|B_j|}{d}\mathbb{E}_t|\nabla_{i_j} f(\mathbf{y}^{(t)})|^2 \\
&= \sum_j \frac{|B_j|}{d}\frac{1}{|B_j|}|\nabla_{B_j} f(\mathbf{y}^{(t)})|^2 \\
&= \frac{1}{d}|\nabla f(\mathbf{y}^{(t)})|^2.
\end{aligned}$$

□

Lemma 1.2. *Define* $\mathbf{t}^{(t+1)} := \mathbf{z}^{(t)} - \frac{1}{n\theta_t} L_{\max}^{-1} \nabla f(\mathbf{y}^{(t)})$. *Or equivalently,*

$$\mathbf{t}^{(t+1)} := \arg\min_{\mathbf{z}} \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{z} - \mathbf{z}^{(t)} \rangle + \frac{d\theta_t L_{\max}}{2} \|\mathbf{z} - \mathbf{z}^{(t)}\|^2.$$

Then

$$\mathbb{E}_t f(\mathbf{x}^{(t+1)}) \leq (1-\theta_t) f(\mathbf{x}^{(t)}) + \theta_t f(\mathbf{x}^*) + \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{t}^{(t+1)}\|^2.$$

Proof. By Lemma 1.1, we have

$$\begin{aligned}
\mathbb{E}_t f(\mathbf{x}^{(t)}) &\leq f(\mathbf{y}^{(t)}) + \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{s}^{(t+1)} - \mathbf{y}^{(t)} \rangle + \frac{dL_{\max}}{2} \|\mathbf{s}^{(t+1)} - \mathbf{y}^{(t)}\|^2 \\
&= f(\mathbf{y}^{(t)}) + \theta_t \big(\langle \nabla f(\mathbf{y}^{(t)}), \mathbf{t}^{(t+1)} - \mathbf{z}^{(t)} \rangle + \frac{dL_{\max}\theta_t}{2} \|\mathbf{t}^{(t+1)} - \mathbf{z}^{(t)}\|^2 \big) \\
&= f(\mathbf{y}^{(t)}) + \theta_t \big(\langle \nabla f(\mathbf{y}^{(t)}), \mathbf{x}^* - \mathbf{z}^{(t)} \rangle + \frac{dL_{\max}\theta_t}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{dL_{\max}L\theta_t}{2} \|\mathbf{x}^* -^{(\mathtt{t+1})} \| \big) \\
&= (1-\theta_t)(f(\mathbf{y}^{(t)}) + \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{x}^{(t)} - \mathbf{y}^{(t)} \rangle) + \theta_t (f(\mathbf{y}^{(t)}) + \langle \nabla f(\mathbf{y}^{(t)}), \mathbf{x}^* - \mathbf{y}^{(t)} \rangle) \\
&\qquad + \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{t}^{(t+1)}\|^2 \\
&\leq (1-\theta_t) f(\mathbf{x}^{(t)}) + \theta_t f(\mathbf{x}^*) + \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{nL_{\max}\theta_t^2}{2} \|\mathbf{x}^* - \mathbf{t}^{(t+1)}\|^2.
\end{aligned}$$

□

Lemma 1.3. $\frac{dL_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{dL_{\max}}{2} \|\mathbf{x}^* - \mathbf{t}^{(t+1)}\|^2 = \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{d^2 L_{\max}}{2} \mathbb{E}_{j_2} \|\mathbf{x}^* - \mathbf{z}^{(t+1)}\|^2.$

Proof.

$$\begin{aligned}
\frac{dL_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{dL_{\max}}{2} \|\mathbf{x}^* - \mathbf{t}^{(t+1)}\|^2 &= \frac{dL_{\max}}{2} \langle \mathbf{t}^{(t+1)} - \mathbf{z}^{(t)}, 2\mathbf{x}^* - 2\mathbf{z}^{(t)} \rangle - \frac{dL_{\max}}{2} \|\mathbf{t}^{(t+1)} - \mathbf{z}^{(t)}\|^2 \\
&= \frac{d^2 L_{\max}}{2} \mathbb{E}_{j_2} [\langle \mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}, 2\mathbf{x}^* - 2\mathbf{z}^{(t)} \rangle - \|\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}\|^2] \qquad (1.20) \\
&= \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{d^2 L_{\max}}{2} \mathbb{E}_{j_2} \|\mathbf{x}^* - \mathbf{z}^{(t+1)}\|^2. \qquad (1.21)
\end{aligned}$$

Here, we use the fact that $\mathbf{t}^{(t+1)} - \mathbf{z}^{(t)} = d\mathbb{E}_{j_2}[\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}]$ and $\|\mathbf{t}^{(t+1)} - \mathbf{z}^{(t)}\|^2 = d\mathbb{E}_{j_2} \|\mathbf{z}^{(t+1)} - \mathbf{z}^{(t)}\|^2$. □

Proof of Theorem 6.11. By Lemma 1.2 and Lemma 1.3, we obtain that

$$\mathbb{E}_t f(\mathbf{x}^{(t+1)}) \leq (1-\theta_t) f(\mathbf{x}^{(t)}) + \theta_t f(\mathbf{x}^*) + \frac{d^2\theta_t^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t)}\|^2 - \frac{d^2\theta_t^2 L_{\max}}{2} \mathbb{E}_{j_2} \|\mathbf{x}^* - \mathbf{z}^{(t+1)}\|^2.$$

By using $\frac{1-\theta_{t+1}}{\theta_{t+1}^2} = \frac{1}{\theta_t^2}$, we arrive at:

$$\frac{1-\theta_{t+1}}{\theta_{t+1}^2} (\mathbb{E}_t f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*)) + \frac{d^2 L_{\max}}{2} \mathbb{E}_{j_2} \|\mathbf{x}^* - \mathbf{z}^{(t+1)}\|^2 \leq \frac{1-\theta_t}{\theta_t^2} (f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^t\|^2$$

We use $\mathbb{E}^t$ to denote taking expectation over everything up to t, it follows that

$$\mathbb{E}^{t+1} \left[\frac{1-\theta_{t+1}}{\theta_{t+1}^2} (f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*)) + \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^{(t+1)}\|^2 \right] \leq \mathbb{E}^t \left[\frac{1-\theta_t}{\theta_t^2} (f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^t\|^2 \right].$$

By above recursive formula, we get

$$\begin{aligned}
\mathbb{E}^{t+1} \left[\frac{1-\theta_{t+1}}{\theta_{t+1}^2} (f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^*)) \right] &\leq \mathbb{E}^0 \left[\frac{1-\theta_0}{\theta_0^2} (f(\mathbf{x}^0) - f(\mathbf{x}^*)) + \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{z}^0\|^2 \right] \qquad (1.22) \\
&= \frac{d^2 L_{\max}}{2} \|\mathbf{x}^* - \mathbf{x}^0\|^2 \qquad (1.23)
\end{aligned}$$

Algorithm 1 Proximal hybrid coordinate descent

Input: $\mathbf{x}^{(0)}, \mathcal{B} = \{B_i\}_{i=1}^k$.
for $t = 0, 1, 2, \ldots$ **do**
  $I = \emptyset$
  for $j = 1, 2, \ldots, k$ **do**
    [**Random rule**] uniform randomly choose a $i_j \in B_j$ and let $I = I \cup \{i_j\}$
  end for
  [**GS-s rule**] $i \in \arg\max_{j \in I} \{\min_{s \in g_i} |\nabla_j f(\mathbf{x}^t) + s|\}$
  $\mathbf{x}^{(t+1)} = \text{prox}_{1/L_i g_i}\left(\mathbf{x}^{(t)} - \frac{1}{L_i}\nabla f_i(\mathbf{x}^{(t)})\mathbf{e}_i\right)$
end for

It is easy to check that $\theta_t \leq \frac{2}{t+2}$, then it gives

$$\mathbb{E}^t\left[f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)\right] \leq \frac{\theta_t^2}{1-\theta_t}\frac{d^2 L_{\max}}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2 = \frac{d^2\theta_{t-1}^2 L_{\max}}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2 \leq \frac{2d^2 L_{\max}}{(t+1)^2}\|\mathbf{x}^* - \mathbf{x}^0\|^2.$$

□

High probability error bounds

The following high probability error bounds can be obtained by using (Richtárik and Takác, 2014, Theorem 1)

Corollary 1.4. *Denote $\mathbf{x}^{(t)}$ as the iterate generated from Algorithm 2. For f that is μ_1 and μ_2 strongly convex with respect for 1 and 2-norm. Let*

$$\eta := \inf_{\mathbf{x}\in\mathbb{R}^d} \max\left\{\frac{\mu_2}{L_{\max}}\frac{\|\nabla f(\mathbf{x})\|_{\mathcal{B},\infty}^2}{\|\nabla f(\mathbf{x})\|_2^2}, \frac{\mu_1}{L_{\max}}\frac{\|\nabla f(\mathbf{x})\|_{\mathcal{B},\infty}^2}{\|\nabla f(\mathbf{x})\|_\infty^2}\right\},$$

then with probability at least $1-\beta$, we have

$$\mathbb{E}[f(\mathbf{x}^{(t)})] - f^* \leq \frac{\exp(-t\eta)}{\beta}(f(\mathbf{x}^0) - f^*).$$

Using Equation (1.5) and (Richtárik and Takác, 2014, Theorem 1), we can immediately get the following.

Corollary 1.5. *Denote $\mathbf{x}^{(t)}$ as the iterate generated from Algorithm 2. For convex objective f, with probability at least $1-\beta$,*

$$\mathbb{E}[f(\mathbf{x}^{(t)})] - f^* = \mathcal{O}\left(\frac{L_{\max}D^2}{\eta^2 t}\left(1 + \log\left(\frac{1}{\beta}\right)\right)\right),$$

where $\rho := \inf_{\mathbf{x}\in\mathbb{R}^d}\{\|\nabla f(\mathbf{x})\|_{\mathcal{B},\infty}^2 / \|\nabla f(\mathbf{x})\|_\infty^2\}$ and $D = \sup_{\mathbf{x}\in\mathbb{R}^d}\{\|\mathbf{x} - \mathbf{x}^*\|_1 \mid f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\}$.

2 PROXIMAL HYBRIDCD

Proximal hybridCD is a proximal-gradient variant of hybridCD. It aims to solve the composite problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(x) + \sum_{i=1}^d g_i(\mathbf{x}_i).$$

The detailed algorithm is shown in Algorithm 1, where

$$\text{prox}_g(\mathbf{y}) := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 + g(\mathbf{y})$$

is the standard definition of proximal operator and the GS-s rule is the greedy selection rule extended to composite problem, see Nutini et al. (2015) for more details.

Bibliography

Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Nearest neighbor based greedy coordinate descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2160–2168, Granada, Spain, 2011.

Sai Praneeth Karimireddy, Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Efficient greedy coordinate descent for composite problems. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2887–2896, Naha, Okinawa, Japan, 2019.

Haihao Lu, Robert M. Freund, and Vahab S. Mirrokni. Accelerating greedy coordinate descent methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3263–3272, Stockholmsmässan, Stockholm, Sweden, 2018.

Julie Nutini, Mark Schmidt, Issam H. Laradji, Michael P. Friedlander, and Hoyt A. Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1632–1641, Lille, France, 2015.

Peter Richtárik and Martin Takác. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.*, 144(1-2):1–38, 2014.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.