
Supplementary Material for Parallel Successive Convex Approximation for Nonsmooth Nonconvex Optimization

Meisam Razaviyayn*
meisamr@stanford.edu

Mingyi Hong†
mingyi@iastate.edu

Zhi-Quan Luo‡
luozq@umn.edu

Jong-Shi Pang§
jongship@usc.edu

1 Proofs

Proof of Equation (5):

Let us first show the result for the randomized block selection rule. We will do so by proving that $\lim_{r \rightarrow \infty} \|\hat{x}^r - x^r\| = 0$, with probability one. To show this, we start by bounding the decrease in the objective value in the consecutive steps of the algorithm:

$$\begin{aligned}
 h(x^{r+1}) &= f(x^{r+1}) + \sum_i g_i(x_i^{r+1}) = f(x^{r+1}) + \sum_{i \notin S^r} g_i(x_i^r) + \sum_{i \in S^r} g_i(x_i^r + \gamma^r(\hat{x}_i^r - x_i^r)) \\
 &\leq f(x^{r+1}) + \sum_i g_i(x_i^r) + \gamma^r \sum_{i \in S^r} (g_i(\hat{x}_i^r) - g_i(x_i^r)) \\
 &\leq f(x^r) + \gamma^r \langle \nabla f(x^r), \hat{x}^r - x^r \rangle_{S^r} + \frac{(\gamma^r)^2 L_{\nabla F}}{2} \|\hat{x}^r - x^r\|_{S^r}^2 + \sum_i g_i(x_i^r) + \gamma^r \sum_{i \in S^r} (g_i(\hat{x}_i^r) - g_i(x_i^r)) \\
 &= h(x^r) + \frac{(\gamma^r)^2 L_{\nabla f}}{2} \|\hat{x}^r - x^r\|_{S^r}^2 + \gamma^r \left(\langle \nabla f(x^r), \hat{x}^r - x^r \rangle_{S^r} + \sum_{i \in S^r} (g_i(\hat{x}_i^r) - g_i(x_i^r)) \right), \tag{1}
 \end{aligned}$$

where the first inequality is due to convexity of $g(\cdot)$; the second inequality is due to the Lipschitz continuity of $\nabla f(\cdot)$; and we have also used the notation $\langle a, b \rangle_S \triangleq \sum_{i \in S} \langle a_i, b_i \rangle$ and $\|a\|_S^2 \triangleq \langle a, a \rangle_S$. In order to get a standard form sufficient decrease bound, we need to bound the last term in (1). Noticing that \tilde{h}_i is strongly convex, the definition of \hat{x}_i^r leads to

$$\tilde{h}_i(x_i^r, x^r) \geq \tilde{h}_i(\hat{x}_i^r, x^r) + \frac{\tau}{2} \|\hat{x}_i^r - x_i^r\|^2, \quad \forall i \in S^r.$$

Substituting the definition of \tilde{h}_i and multiplying both sides by minus one give

$$-\tilde{f}_i(x_i^r, x^r) - g_i(x_i^r) \leq -\tilde{f}_i(\hat{x}_i^r, x^r) - g_i(\hat{x}_i^r) - \frac{\tau}{2} \|\hat{x}_i^r - x_i^r\|^2.$$

Linearizing the smooth part, the gradient consistency assumption leads to

$$\langle \nabla_{x_i} f(x^r), \hat{x}_i^r - x_i^r \rangle + g_i(\hat{x}_i^r) - g_i(x_i^r) \leq -\frac{\tau}{2} \|\hat{x}_i^r - x_i^r\|^2.$$

Summing up the above inequality over all $i \in S^r$, we obtain

$$\langle \nabla f(x^r), \hat{x}^r - x^r \rangle_{S^r} + \sum_{i \in S^r} (g_i(\hat{x}_i^r) - g_i(x_i^r)) \leq -\frac{\tau}{2} \|\hat{x}^r - x^r\|_{S^r}^2, \tag{2}$$

*Electrical Engineering Department, Stanford University

†Industrial and Manufacturing Systems Engineering, Iowa State University

‡Department of Electrical and Computer Engineering, University of Minnesota

§Department of Industrial and Systems Engineering, University of Southern California

where $\hat{x}^r \triangleq (\hat{x}_i^r)_{i=1}^n$. Combining (1) and (2) leads to

$$h(x^{r+1}) \leq h(x^r) + \frac{\gamma^r(-\tau + \gamma^r L_{\nabla f})}{2} \|\hat{x}^r - x^r\|_{S^r}^2.$$

Proof of Lemma 3:

Let us first prove the cost-to-go bound for the randomized case. It can be observed that the conditional expected cost-to-go can be bounded by

$$\begin{aligned} \mathbb{E} [h(x^{r+1}) - h(x^*) \mid x^r] &\stackrel{(i)}{\leq} h(x^r) - h(x^*) = f(x^r) - f(x^*) + g(x^r) - g(x^*) \\ &\stackrel{(ii)}{\leq} \langle \nabla f(x^r), x^r - \hat{x}^r \rangle + \langle \nabla f(x^r), \hat{x}^r - x^* \rangle + L_g \|x^r - \hat{x}^r\| + g(\hat{x}^r) - g(x^*) \\ &\stackrel{(iii)}{\leq} (L_g + Q) \|\hat{x}^r - x^r\| + \sum_{i=1}^n \langle \nabla_{x_i} f(x^r) - \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r) + \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r), \hat{x}_i^r - x_i^* \rangle + g(\hat{x}^r) - g(x^*) \\ &\leq (L_g + Q) \|\hat{x}^r - x^r\| + \sum_{i=1}^n \langle \nabla_{x_i} f(x^r) - \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r), \hat{x}_i^r - x_i^* \rangle \end{aligned} \quad (3)$$

where (i) is due to the Sufficient Descent Lemma; the inequality (ii) is due to the convexity of $f(\cdot)$ and Lipschitz continuity of $g(\cdot)$; the third inequality is by the bounded level set assumption. Furthermore, the last inequality is obtained by exploiting the first order optimality condition of the point \hat{x}_i^r , i.e., $\langle \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r), \hat{x}_i^r - x_i^* \rangle + g_i(\hat{x}_i^r) - g_i(x_i^*) \leq 0$. In addition to the above inequality, one can easily deduce

$$\begin{aligned} \left(\sum_{i=1}^n \langle \nabla_{x_i} f(x^r) - \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r), \hat{x}_i^r - x_i^* \rangle \right)^2 &= \left(\sum_{i=1}^n \langle \nabla_{x_i} \tilde{f}_i(x_i^r, x^r) - \nabla_{x_i} \tilde{f}_i(\hat{x}_i^r, x^r), \hat{x}_i^r - x_i^* \rangle \right)^2 \\ &\leq n \sum_{i=1}^n L_i^2 \|x_i^r - \hat{x}_i^r\|^2 \cdot \|\hat{x}_i^r - x_i^*\|^2 \leq n L^2 R^2 \|x^r - \hat{x}^r\|^2. \end{aligned} \quad (4)$$

Combining (3) and (4) will conclude the proof for the randomized case.

For the cyclic case, we can simply bound the cost-to-go estimate by

$$\begin{aligned} h(x^{m(r+1)}) - h(x^*) &= f(x^{m(r+1)}) - f(x^*) + g(x^{m(r+1)}) - g(x^*) \\ &\leq \langle \nabla f(x^{m(r+1)}), x^{m(r+1)} - x^* \rangle + g(x^{m(r+1)}) - g(x^*) \end{aligned} \quad (5)$$

$$\begin{aligned} &= \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i f(x^{m(r+1)}), x_i^{m(r+1)} - x_i^* \rangle + g_i(x^{m(r+1)}) - g_i(x^*) \\ &\leq \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i f(x^{m(r+1)}) - \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), x_i^{m(r+1)} - x_i^* \rangle \end{aligned} \quad (6)$$

$$+ \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), x_i^{m(r+1)} - \hat{x}_i^{mr+\ell} \rangle \quad (7)$$

$$+ \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), \hat{x}_i^{mr+\ell} - x_i^* \rangle + g_i(\hat{x}_i^{mr+\ell}) - g_i(x_i^*) \quad (8)$$

$$+ \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} g_i(x_i^{m(r+1)}) - g_i(\hat{x}_i^{mr+\ell}), \quad (9)$$

where (5) is due to convexity of the function $f(\bullet)$. First notice that (8) is nonpositive due to the definition of $\hat{x}_i^{mr+\ell}$ and the optimality condition for it. Now we bound the terms (6), (7), and (9) separately. Let us first start by bounding the terms in (6):

$$\begin{aligned}
& \left(\sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i f(x^{m(r+1)}) - \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), x_i^{m(r+1)} - x_i^* \rangle \right)^2 \\
& \leq n \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} R^2 \|\nabla_i f(x^{m(r+1)}) - \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell})\|^2 \\
& = nR^2 \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \|\nabla_i \tilde{f}_i(x_i^{m(r+1)}, x^{m(r+1)}) - \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell})\|^2 \\
& \leq 2nR^2 \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \left(\|\nabla_i \tilde{f}_i(x_i^{m(r+1)}, x^{m(r+1)}) - \nabla_i \tilde{f}_i(x_i^{m(r+1)}, x^{mr+\ell})\|^2 \right. \\
& \quad \left. + \|\nabla_i \tilde{f}_i(x_i^{m(r+1)}, x^{mr+\ell}) - \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell})\|^2 \right) \\
& \leq 2nR^2 \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \left(\tilde{L}^2 \|x^{m(r+1)} - x^{mr+\ell}\|^2 + L_i^2 \|x_i^{m(r+1)} - \hat{x}_i^{mr+\ell}\|^2 \right) \\
& \leq 2nR^2 \left(n\tilde{L}^2 \|x^{m(r+1)} - x^{mr}\|^2 + L^2 \frac{(1-\gamma)^2}{\gamma^2} \|x^{m(r+1)} - x^{mr}\|^2 \right) \\
& = 2nR^2 \left(n\tilde{L}^2 + \frac{L^2(1-\gamma)^2}{\gamma^2} \right) \|x^{m(r+1)} - x^{mr}\|^2. \tag{10}
\end{aligned}$$

Now, we can bound (7) by

$$\begin{aligned}
& \left(\sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \langle \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), x_i^{m(r+1)} - \hat{x}_i^{mr+\ell} \rangle \right)^2 \\
& \leq n \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \left(\langle \nabla_i \tilde{f}_i(\hat{x}_i^{mr+\ell}, x^{mr+\ell}), x_i^{m(r+1)} - \hat{x}_i^{mr+\ell} \rangle \right)^2 \\
& \leq n \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \hat{Q}^2 \|x_i^{m(r+1)} - \hat{x}_i^{mr+\ell}\|^2 \\
& = n\hat{Q}^2 \sum_i \frac{(1-\gamma)^2}{\gamma^2} \|x_i^{m(r+1)} - x_i^{mr}\|^2 \\
& = n\hat{Q}^2 \frac{(1-\gamma)^2}{\gamma^2} \|x^{m(r+1)} - x^{mr}\|^2. \tag{11}
\end{aligned}$$

Finally, we can bound (9) by

$$\begin{aligned}
& \left(\sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} g_i(x_i^{m(r+1)}) - g_i(\hat{x}_i^{mr+\ell}) \right)^2 \leq n \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} \left(g_i(x_i^{m(r+1)}) - g_i(\hat{x}_i^{mr+\ell}) \right)^2 \\
& \leq n \sum_{\ell=0}^{m-1} \sum_{i \in \mathcal{T}_\ell} L_g^2 \|x_i^{m(r+1)} - \hat{x}_i^{mr+\ell}\|^2 \\
& = nL_g^2 \sum_i \frac{(1-\gamma)^2}{\gamma^2} \|x_i^{m(r+1)} - x_i^{mr}\|^2 \\
& = nL_g^2 \frac{(1-\gamma)^2}{\gamma^2} \|x^{m(r+1)} - x^{mr}\|^2. \tag{12}
\end{aligned}$$

Plugging (10), (11), (12) in (6), (7), (9) implies the desired cost-to-go bound.

2 Additional Numerical Experiments

In this section, we present two additional numerical experiments. Similar to the main body of the manuscript, we consider LASSO problem and the data is generated as explained before. Here we only consider the large size experiment, i.e., $A \in \mathbb{R}^{10,000 \times 100,000}$. In the first simulation, we see the effect of changing the parameters of the algorithm. As discussed in the paper, too large or too small step-size could result in slower convergence speed.

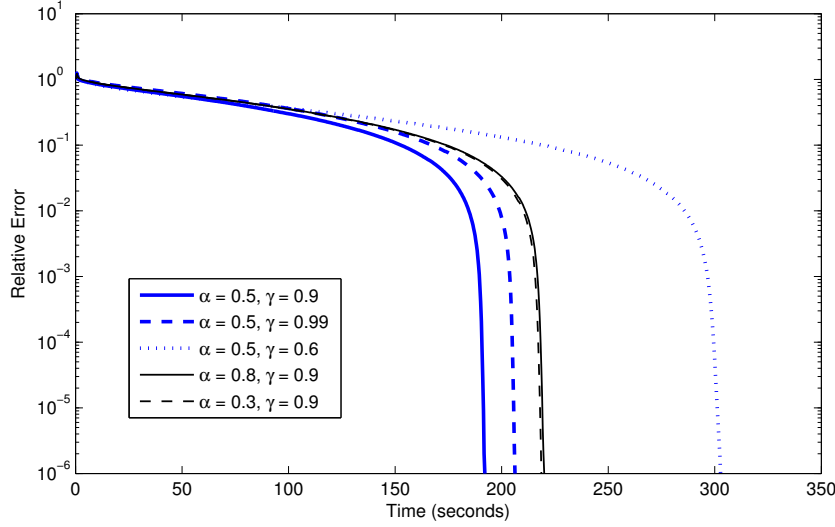


Figure 1: Performance of the algorithm with different choices of α and γ

As we saw in the main body of the manuscript, the overall convergence time of the algorithm does not always improve as the number of processors increases. This fact is due to the communication overhead among the processing nodes. In Figure 2, we only plot the computation time and ignore the communication time. As can be seen in this plot, the computation time spent on the nodes always decreases by utilizing more processors.

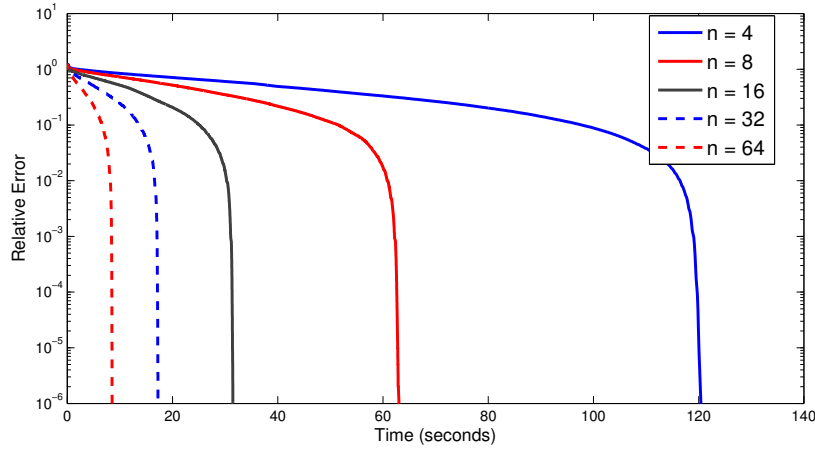


Figure 2: Computation time of the algorithm for different number of processors