Distributed Subgradient Descent

Mingyi Hong

University Of Minnesota

Outline

- The DGD/DSD Algorithm
- Assumptions and Convergence
- Convergence Rates
- Summary: Benefit/Disadvantage of DGD/DSD

Unconstrained multiagent-optimization problem

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & \sum_{i=1}^{m} f_i(x) \\ \text{subject to} \quad & x \in \mathbb{R}^n \end{aligned} \tag{1.1}$$

- Each $f_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function, representing the local-objective function of agent i, which is known only to this agent.
- We do not assume differentiability of any of the functions f_i . At the points where the function fails to be differentiable, a subgradient exists.

Distributed Subgradient Descent (DSG/DGD)

- Specifically, agent i updates its estimates by setting

$$x_i(k+1) = \underbrace{\sum_{j=1}^{m} w_{ij}(k) x_j(k)}_{\text{Consensus Step}} - \underbrace{\alpha \times d_i(k)}_{\text{Subgradient Step}} \tag{1.2}$$

  where the scalar $\alpha > 0$ is a stepsize and the vector $d_i(k)$ is a subgradient of the agent i cost function $f_i(x)$ at $x = x_i(k)$.

- The scalars $w_{i1}(k), \cdots, w_{im}(k)$ are non-negative weights that agent i gives to the estimates $x_1(k), \cdots, x_m(k)$.

Distributed Subgradient Descent (DSD/DGD)

- We could also write the update rule (1.2) in matrix-form

$$X(k+1) = W(k)X(k) - \alpha D(k) \tag{1.3}$$

where $X(k) = \begin{pmatrix} x_1(k)^T \\ x_2(k)^T \\ \vdots \\ x_m(k)^T \end{pmatrix}$, $D(k) = \begin{pmatrix} d_1(k)^T \\ d_2(k)^T \\ \vdots \\ d_m(k)^T \end{pmatrix}$

and

$$W(k) = [w_{ij}(k)]_{i,j=1,\cdots,m}.$$

Note, that the notation here is slightly different than the previous lecture (assume x_i's are vectors now)

Intuition for DGD

- Before proceed, let us understand the advantage / disadvantage of DGD/DSG
- Assume the mixing matrix is fixed, $W(k) = W \ \forall k$, and given the update of DGD as

$$X(k+1) = WX(k) - \alpha D(k),$$

- **Question:** Whether DGD can converge to the global minimum of the objective (1.1)?

Intuition for DGD

- When we set $k \to \infty$, it becomes

$$X(\infty) = WX(\infty) - \alpha D(\infty)$$

- Assuming DGD could achieve consensus such that

$$X(\infty) = WX(\infty), x_i(\infty) = x_j(\infty) = x^* \text{ for all i, j}$$

  Then we get $D(\infty) = 0$ which means $d_i(x^*) = 0$ for all i.

- However, it is impossible for the same point x^* simultaneously minimizes f_i to achieve $d_i(x^*) = 0$ for all i!

Convergence analysis of the subgradient method

- The previous discussion implies that, if a constant stepsize α is used ([that is, independent of iteration number k]), then DGD/DSD will *not* converge to global min

- Now we study the convergence properties of the DGD/DSG method (1.2)

Assumption

- **Assumption 1** : For any time $k \geq 0$, the weight matrix $W(k)$ is doubly stochastic with positive diagonal. Additionally, there is a scalar $\eta > 0$ such that if $w_{ij}(k) > 0$, then $w_{ij}(k) \geq \eta$.

Assumption

- **Assumption 2**: There exists an integer $B \geq 1$ such that the directed graph

$$(\mathcal{N}, \xi(W(kB)) \cup \cdots \cup \xi(W((k+1)B-1)))$$

is strongly connected for all $k \geq 0$, where the link set $\xi(W(k))$ is given by

$$\xi(W(k)) = \{(j,i) | a_{ij}(k) > 0, i, j = 1, \cdots, m\} \text{ for all } k.$$

Assumption

- **Assumption 3**: Assume the uniform boundedness of the set of subgradients of the cost functions f_i at all points: for some scalar $L > 0$, we have for all $x \in \mathbb{R}^n$ and all i,

$$\|g\| \leq L \quad \text{for all } g \in \partial f_i(x),$$

where $\partial f_i(x)$ is the set of all subgradients of f_i at x.

Convergence analysis of the subgradient method

- The analysis combines the proof techniques used for consensus algorithms and approximate subgradient methods.

- The consensus analysis rests on the convergence rate result of Theorem 1 in our last lecture for transition matrices, which provides a tool for measuring the "agent disagreements" $\|x_i(k) - x_j(k)\|$

The main steps

- Equivalently, we can measure $\|x_i(k) - x_j(k)\|$ in terms of the disagreements $\|x_i(k) - y(k)\|$ with respect to an auxiliary sequence $\{y(k)\}$, defined appropriately (as an average of the iterates).

- The sequence y_k will also serve as a basis for understanding the effects of subgradient steps in the algorithm.

- Proof Technique:
  1. establish the suboptimality of the "average" sequence $y(k)$;
  2. using the estimates for the disagreements $\|x_i(k) - y(k)\|$;
  3. provide a performance bound for the algorithm

The auxiliary variable

- To estimate the agent "disagreements", we use an auxiliary sequence $\{y(k)\}$ of reference points, defined as follows:

$$y(k+1) = y(k) - \frac{\alpha}{m} \sum_{i=1}^{m} d_i(k) \tag{1.4}$$

where $d_i(k)$ is the same subgradient of $f_i(x)$ at $x = x_i(k)$ used in DGD, and

$$y(0) = \frac{1}{m} \sum_{i=1}^{m} x_i(0).$$

The auxiliary variable

- From the definition of the sequence $\{y(k)\}$ in (1.4), it follows for all k

$$y(k) = \frac{1}{m}\sum_{i=1}^{m} x_i(0) - \frac{\alpha}{m}\sum_{r=0}^{k-1}\sum_{i=1}^{m} d_i(r) \tag{1.5}$$

- What is the connection between $X(k)$ and the anuiliary $\{y(k)\}$?

The auxiliary variable

- Given the doubly-stochastic matrix $A(k)$ and the DGD update rule (1.3), we denote the averge of $X(t)$ as $\bar{x}(k)$ as follows

$$\bar{x}(k) = \frac{1}{m} 1^T X(k) = \frac{1}{m} 1^T \left(AX(k-1) - \alpha D(k-1) \right)$$

Therefore, we obtain that

$$\bar{x}(k) = \frac{1}{m} 1^T X(k-1) - \frac{\alpha}{m} 1^T D(k-1) = \bar{x}(k-1) - \frac{\alpha}{m} \sum_{i=1}^{m} d_i(k-1)$$

- We get

$$\bar{x}(k) = \frac{1}{m} \sum_{i=1}^{m} x_i(0) - \frac{\alpha}{m} \sum_{r=0}^{k-1} \sum_{i=1}^{m} d_i(r) = y(k).$$

Preliminary Results

Let's recall the convergence results we established last time.

Lemma 2.1

Let Assumption 1-2 hold, it can be shown that $V(k)$ is nonincreasing in k. Furthermore,

$$V((k+1)B) \leq \left(1 - \frac{\eta}{2m^2}\right) V(kB) \text{ for all } k \geq 0.$$

Theorem 2.2

Let Assumption 1-2 hold, for all i, j and all k, s with $k \geq s$, we have

$$\left| [\Phi(k,s)]_{ij} - \frac{1}{m} \right| \leq (1 - \frac{\eta}{4m^2})^{\lceil \frac{k-s+1}{B} \rceil - 2}.$$

Bound the iterates and the average

In the following lemma, we estimate the norms of the differences $x_i(k) - y(k)$ at each time k. The result relies on Theorem 2.2.

Lemma 2.3

Let Assumption 1-3 hold. Then for all i and $k \geq 1$,

$$\|x_i(k) - y(k)\| \leq \beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| + \alpha L \left(2 + \frac{mB}{\beta(1-\beta)}\right),$$

where $\beta = 1 - \frac{\eta}{4m^2}$.

Proof of Lemma 2.3

Proof of Lemma 2.3.

From the definition of the sequence $\{y(k)\}$ in (1.4), it follows for all k

$$y(k) = \frac{1}{m}\sum_{i=1}^{m} x_i(0) - \frac{\alpha}{m}\sum_{r=0}^{k-1}\sum_{i=1}^{m} d_i(r)$$

Also from the transition matrices $\Phi(k,s)$, we have for all k

$$x_i(k+1) = \sum_{j=1}^{m}[\Phi(k,s)]_{ij}x_j(s) - \alpha\sum_{r=s}^{k-1}\sum_{j=1}^{m}[\Phi(k,r+1)]_{ij}d_j(r) - \alpha d_i(k).$$

From the equation above (set $s=0$), we further get

$$x_i(k) = \sum_{j=1}^{m}[\Phi(k-1,0)]_{ij}x_j(0) - \alpha\sum_{r=0}^{k-2}\sum_{j=1}^{m}[\Phi(k-1,r+1)]_{ij}d_j(r) - \alpha d_i(k-1)$$

Proof of Lemma 2.3

By subtracting the previous two equations, we obtain for all $k \geq 1$,

$$\begin{aligned} x_i(k) - y(k) = & \sum_{j=1}^{m} \left([\Phi(k-1,0)]_{ij} - \frac{1}{m} \right) x_j(0) \\ & - \alpha \sum_{r=0}^{k-2} \sum_{j=1}^{m} \left([\Phi(k-1,r+1)]_{ij} - \frac{1}{m} \right) d_j(r) \\ & - \alpha d_i(k-1) + \frac{\alpha}{m} \sum_{i=1}^{m} d_i(k-1) \end{aligned}$$

Proof of Lemma 2.3

Further utilizing the subgradient boundedness assumption (11), we obtain for all $k \geq 1$

$$\|x_i(k) - y(k)\| \leq \sum_{j=1}^{m} \left| [\Phi(k-1,0)]_{ij} - \frac{1}{m} \right| \times \|x_j(0)\| + \alpha L \sum_{s=1}^{k-1} \sum_{j=1}^{m} \left| [\Phi(k-1,s)]_{ij} - \frac{1}{m} \right| + 2\alpha L$$

Next, recall the convergence theorem of transition matrices, we could further bound the terms $\left| [\Phi(k-1, r+1)]_{ij} - \frac{1}{m} \right|$.

Proof of Lemma 2.3

By Theorem 2.2, we obtain for all i and any $k \geq 1$,

$$\|x_i(k) - y(k)\| \leq \beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| + \alpha L \sum_{s=1}^{k-1} \sum_{j=1}^{m} \beta^{\lceil \frac{k-s}{B} \rceil - 2} + 2\alpha L$$

$$= \beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| + \alpha L m \sum_{s=1}^{k-1} \beta^{\lceil \frac{k-s}{B} \rceil - 2} + 2\alpha L$$

Finally, we could bound the second term above to finish this proof.

Convergence analysis of the subgradient method

Note that

$$\sum_{s=1}^{k-1} \beta^{\lceil \frac{k-s}{B} \rceil - 2} \leq \sum_{r=1}^{\infty} \beta^{\lceil \frac{r}{B} \rceil - 2} = \frac{1}{\beta} \sum_{r=1}^{\infty} \beta^{\lceil \frac{r}{B} \rceil - 1}$$

$$\sum_{r=1}^{\infty} \beta^{\lceil \frac{r}{B} \rceil - 1} \leq B \sum_{t=0}^{\infty} \beta^{t} = \frac{B}{1-\beta}.$$

We obtain

$$\sum_{s=1}^{k-1} \beta^{\lceil \frac{k-s}{B} \rceil - 2} \leq \frac{B}{\beta(1-\beta)}.$$

Finally, we get the bound that for all $k \geq 1$,

$$\|x_i(k) - y(k)\| \leq \beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| + \alpha L \left(2 + \frac{mB}{\beta(1-\beta)}\right).$$

Functional Value Convergence

- We next establish a result that estimates the objective function $f = \sum_{i=1}^{m} f_i$ at the running averages of the vectors $y(k)$ of (1.5).
- Specifically, we define $\hat{y}(k) = \frac{1}{k} \sum_{h=1}^{k} y(k)$ for all $k \geq 1$, and we estimate the function values $f(\hat{y}(k))$.
- Through establishing a result to estimate the function values $f(\hat{y}(k))$, we can finally link $\hat{y}(k)$ to the optimization variable x. Finally, we could establish a performance bound on x.

Key Steps

Several key step to finish the following Lemma:

- In order to establish the performance bound of the objective value $f(\hat{y}(k)) - f^*$, we need to expand $\|y(k+1) - x^*\|^2$.

- Then through using the subgradient boundness assumption or some Lipschitz conditions, we are able to convert the distance norm to objective function value.

- Finally, we can combine previous constructed lemma and finish the proof.

Bounding the Functional Value on The Average

Lemma 2.4

Let Assumption 1-3 hold. Also, assume that the set X^* of optimal solutions of the problem (1.1) is nonempty. Then, the average vectors $\hat{y}(k)$ satisfy for all $k \geq 1$,

$$f(\hat{y}(k)) \leq f^* + \frac{\alpha L^2 C}{2} + \frac{2mLB}{k\beta(1-\beta)} \sum_{j=1}^{m} \|x_j(0)\| + \frac{m}{2\alpha k}(dist(y(0), X^*) + 2\alpha L)^2,$$

where $y(0) = \frac{1}{m}\sum_{j=1}^{m} x_j(0)$, $\beta = 1 - \frac{\eta}{4m^2}$ and $C = 1 + 4m\left(2 + \frac{mb}{\beta(1-\beta)}\right)$.

The Next Step

- We establish a bound on the performance of the algorithm at the time-average of the vectors $x(k)$ generated by DGD (1.2).
- We define $\hat{x}_i(k) := \frac{1}{k}\sum_{h=1}^{k} x_i(h)$. Then we bound the objective function improvement at every iteration.
- We will not discuss this step in too much detail

Bounding the functional value on local variables

Theorem 2.5

Let Assumptions 1-3 hold, and assume that the set X^* of optimal solutions of problem (1.1) is nonempty. Then, the averages $\hat{x}_i(k)$ of the iterates obtained by DGD (1.2) satisfy for all i and $k \geq 1$,

$$f(\hat{x}_i(k)) \leq f^* + \frac{\alpha L^2 C_1}{2} + \frac{4mLB}{k\beta(1-\beta)} \sum_{j=1}^{m} \|x_j(0)\| + \frac{m}{2\alpha k}(\mathit{dist}(y(0), X^*) + \alpha L)^2,$$

where $y(0) = \frac{1}{m}\sum_{j=1}^{m} x_j(0)$, $\beta = 1 - \frac{\eta}{4m^2}$ and

$$C_1 = 1 + 8m\left(2 + \frac{mB}{\beta(1-\beta)}\right).$$

Discussion on DGD

- In Lemma 2.3, we see that
  1. $y(k) = \frac{1}{m}\sum_{i=1}^{m} x_i(0) - \frac{\alpha}{m}\sum_{r=0}^{k-1}\sum_{i=1}^{m} d_i(r)$
  2. $\|x_i(k) - y(k)\| \leq \beta^{\lceil \frac{k}{B} \rceil - 2}\sum_{j=1}^{m}\|x_j(0)\| + \underbrace{\alpha L\left(2 + \frac{mB}{\beta(1-\beta)}\right)}_{\text{Constant Gap}}$

- According to this Lemma, it shows that given a constant stepsize, distributed subgradient decent (DGD) do not converge to a solution x^* to achieve the consensus.

- What DGD actually converge to is a point in its neighborhood no matter whether f_i are differentiable or not.

Discussion on DGD

- DGD (1.3) do not converge to a solution x^* of the problem (1.1), but a point in its neighborhood.
- This motivates the use of certain diminishing stepsizes in Chen-2012[1], Duchi-2011 [2] to guarantee convergence to a exact solution x^*.
- Diminishing stepsize has slower convergence rate in general.

[1] Chen, Annie I-An. Fast distributed first-order methods. Diss. Massachusetts Institute of Technology, 2012.

[2] Duchi, John C., Alekh Agarwal, and Martin J. Wainwright. "Dual averaging for distributed optimization: Convergence analysis and network scaling." IEEE Transactions on Automatic control 57.3 (2011): 592-606.

Discussion on DGD

- There are many other methods that are able to use constant stepsize, and converges to global min
- Such as the ADMM method and EXTRA method [3]
- This will be the topic of the next lecture

[3] Shi, Wei, et al. "Extra: An exact first-order algorithm for decentralized consensus optimization." SIAM Journal on Optimization 25.2 (2015): 944-966.

Discussion on DGD

- The analysis of DGD algorithm we introduce is to achieve convergence on the convex objective (1.1).
- However, in an earlier work [4], DGD-type algorithm has been proposed for problems without needing the objective to be convex

[4] Tsitsiklis, John, Dimitri Bertsekas, and Michael Athans. "Distributed asynchronous deterministic and stochastic gradient optimization algorithms." IEEE Transactions on Automatic Control 31.9 (1986): 803-812.

Discussion on DGD

- The most important assumptions in Tsitsiklis' work is to assume that the objective function f has Lipschitz gradient such that

$$\|\nabla f(x) - \nabla f(x')\| \leq K\|x - x'\|,$$

  where K is some nonnegative constant.

- The convergence of several settings are analyzed in Tsitsiklis's work, including distributed asynchronous optimization in deterministic and stochastic cases.

Discussion on DGD

- We mention that under reasonable conditions, the algorithms proposed in Tsitsiklis' work can convergence to the global optimum (in convex case) or a stationary point (in non-convex case).

- Comparing the convergence result between DGD algorithm (introduced today) and Tsitsiklis' work, we see that the DGD algorithm we just learned has the convergence rate but the result in Tsitsiklis' work is to achieve convergence asymptotically.

Discussion on DGD

- In the next lecture, we will introduce several algorithm to achieve convergence for non-convex decentralized optimization.

Appendix and Additional Proofs

Proof of Lemma 2.4

Proof of Lemma 2.4:
From the definition of the sequence $y(k)$, it follows for any $x^* \in X^*$ and all k,

$$\begin{aligned}
&\|y(k+1) - x^*\|^2 \\
=&\|y(k) - \frac{\alpha}{m}\sum_{i=1}^{m} d_i(k) - x^*\|^2 \\
=&\|y(k) - x^*\|^2 + \frac{\alpha^2}{m^2}\|\sum_{i=1}^{m} d_i(k)\|^2 - 2\frac{\alpha}{m}\sum_{i=1}^{m} d_i(k)'\,(y(k) - x^*)
\end{aligned} \tag{3.1}$$

Proof of Lemma 2.4

We next estimate the terms $d_i(k)'(y(k) - x^*)$ where $d_i(k)$ is a subgradient of f_i at $x_i(k)$. For any i and k, we have

$$d_i(k)'(y(k) - x^*) = d_i(k)'(y(k) - x_i(k)) + d_i(k)'(x_i(k) - x^*)$$

By the subgradient property, we have (key)

$$d_i(k)'(x_i(k) - x^*) \geq f_i(x_i(k)) - f_i(x^*).$$

Combining the above two, we get

$$\begin{aligned}
& d_i(k)'(y(k) - x^*) \\
\geq & d_i(k)'(y(k) - x_i(k)) + f_i(x_i(k)) - f_i(x^*) \\
\geq & \underbrace{-L\|y(k) - x_i(k)\|}_{\text{subgradient boundedness}} + [f_i(x_i(k)) - f_i(y(k))] + [f_i(y(k)) - f_i(x^*)]
\end{aligned}$$

Proof of Lemma 2.4

We next consider $f_i(x_i(k)) - f_i(y(k))$, by the subgradient property we have

$$f_i(x_i(k)) - f_i(y(k)) \geq \tilde{d}_i(k)' \left(x_i(k) - y(k)\right) \geq -L\|x_i(k) - y(k)\|$$

where $\tilde{d}_i(k)$ is a subgradient of f_i at $y(k)$.
Through combining the preceding two relations, the following holds for all i and k,

$$d_i(k)' \left(y(k) - x^*\right) \geq -2L\|x_i(k) - y(k)\| + f_i(y(k)) - f_i(x^*) \quad (3.2)$$

Proof of Lemma 2.4

Plugging (3.2) into the (3.1), we could obtain

$$\begin{aligned}
&\|y(k+1) - x^*\|^2 \\
&\leq \|y(k) - x^*\|^2 + \frac{\alpha^2}{m^2}\|\sum_{i=1}^{m} d_i(k)\|^2 \\
&\quad + \frac{4L\alpha}{m}\sum_{i=1}^{m}\|y(k) - x_i(k)\| - \frac{2\alpha}{m}\sum_{i=1}^{m}(f_i(y(k)) - f_i(x^*))
\end{aligned}$$

Proof of Lemma 2.4

Then we can write it as

$$\begin{aligned}
&\|y(k+1) - x^*\|^2 \\
\leq &\|y(k) - x^*\|^2 + \frac{\alpha^2 L^2}{m} + \frac{4L\alpha}{m} \sum_{i=1}^{m} \|y(k) - x_i(k)\| \\
&- \frac{2\alpha}{m} \left(f(y(k)) - f^*\right)
\end{aligned} \tag{3.3}$$

where we note that $f = \sum_{i=1}^{m} f_i$ and $f(x^*) = f^*$.

Proof of Lemma 2.4

Taking the minimum over $x^* \in X^*$ in both sides of the equation (3.3) above, we obtain

$$\begin{aligned}\mathsf{dist}^2(y(k+1), X^*) \leq & \mathsf{dist}^2(y(k), X^*) + \frac{\alpha^2 L^2}{m} \\ & + \frac{4L\alpha}{m} \sum_{i=1}^{m} \|y(k) - x_i(k)\| - \frac{2\alpha}{m} \left(f(y(k)) - f^*\right)\end{aligned}$$

Proof of Lemma 2.4

By using Lemma 2.3 to bound each of the terms $\|y(k) - x_i(k)\|$, we further obtain

$$\begin{aligned}\text{dist}^2(y(k+1), X^*) \leq & \text{dist}^2(y(k), X^*) + \frac{\alpha^2 L^2}{m} + 4\alpha L \beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| \\ & + 4\alpha^2 L^2 \left(2 + \frac{mB}{\beta(1-\beta)}\right) - \frac{2\alpha}{m} \left(f(y(k)) - f^*\right)\end{aligned}$$

We regroup the terms and introduce $C = 1 + 4m\left(2 + \frac{mB}{\beta(1-\beta)}\right)$, it is shown that

$$\begin{aligned}f(y(k)) \leq & f^* + \frac{\alpha L^2 C}{2} + 2mL\beta^{\lceil \frac{k}{B} \rceil - 2} \sum_{j=1}^{m} \|x_j(0)\| \\ & + \frac{m}{2\alpha} \left(\text{dist}^2\left(y(k), X^*\right) - \text{dist}^2\left(y(k+1), X^*\right)\right)\end{aligned}$$

Proof of Lemma 2.4

By adding these inequalities for different values of k, we obtain

$$\frac{1}{k}\sum_{h=1}^{k} f(y(k)) \leq f^* + \frac{\alpha L^2 C}{2} + \frac{2mLB}{k\beta(1-\beta)}\sum_{j=1}^{m}\|x_j(0)\| + \frac{m}{2\alpha k}\left(\mathrm{dist}^2\left(y(1), X^*\right) \underbrace{-\mathrm{dist}^2\left(y(k+1), X^*\right)}_{\text{to be discarded}}\right) \tag{3.4}$$

where this inequality follows that for $t \geq 1$,

$$\sum_{k=1}^{t} \beta^{\lceil \frac{k}{B} \rceil - 2} \leq \frac{1}{\beta}\sum_{k=1}^{\infty} \beta^{\lceil \frac{k}{B} \rceil - 1} \leq \frac{1}{\beta}\sum_{s=0}^{\infty} \beta^s = \frac{B}{\beta(1-\beta)}.$$

Proof of Lemma 2.4

By discarding the nonpositive term on the right-hand side in relation (3.4) and by using the convexity of f,

$$f(\hat{y}(k)) \leq \frac{1}{k}\sum_{h=1}^{k} f(y(k)) \leq f^* + \frac{\alpha L^2 C}{2} + \frac{2mLB}{k\beta(1-\beta)}\sum_{j=1}^{m}\|x_j(0)\| + \frac{m}{2\alpha k}\text{dist}^2(y(1), X^*).$$

Moreover, by the definition of $y(k)$ (1.4) and the subgradient boundedness, we see that

$$\text{dist}^2(y(1), X^*) \leq \left(\text{dist}(\text{y}(0), X^*) + \alpha L\right)^2,$$

which when combined with the preceding relation yields the desired inequality. □

Proof of Theorem 2.5

Proof of Theorem 2.5:

By the convexity of the functions f_j, we have for any i and $k \geq 1$,

$$f(\hat{x}_i(k)) \leq f(\hat{y}(k)) + \sum_{j=1}^{m} g_{ij}(k)' \left(\hat{x}_i(k) - \hat{y}(k)\right),$$

where $g_{ij}(k)$ is a subgradient of f_j at $\hat{x}_i(k)$. Then, by using the subgradient boundedness, we obtain for all i and $k \geq 1$,

$$\begin{aligned} f(\hat{x}_i(k)) &\leq f(\hat{y}(k)) + L \sum_{j=1}^{m} \|\hat{x}_i(k) - \hat{y}(k)\| \\ &\leq f(\hat{y}(k)) + L \sum_{j=1}^{m} \| \frac{1}{k} \sum_{t=1}^{k} (x_i(t) - y(t)) \| \\ &\leq f(\hat{y}(k)) + \frac{2L}{k} \sum_{j=1}^{m} \sum_{t=1}^{k} \underbrace{\|x_i(t) - y(t)\|}_{\text{to be bounded}} \end{aligned} \tag{3.5}$$

Proof of Theorem 2.5

Then we utilize the Lemma 2.3 to bound $\|x_i(t) - y(t)\|$, we have for all i and $k \geq 1$,

$$\sum_{t=1}^{k} \|x_i(t) - y(t)\| \leq \underbrace{\left(\sum_{t=1}^{k} \beta^{\lceil \frac{t}{B} \rceil - 2} \right)}_{\text{to be bounded}} \sum_{j=1}^{m} \|x_j(0)\| + \alpha k L \left(2 + \frac{mB}{\beta(1-\beta)} \right)$$

$$\leq \frac{B}{\beta(1-\beta)} \sum_{j=1}^{m} \|x_j(0)\| + \alpha k L \left(2 + \frac{mB}{\beta(1-\beta)} \right) \tag{3.6}$$

since $\sum_{k=1}^{t} \beta^{\lceil \frac{k}{B} \rceil - 2} \leq \frac{1}{\beta} \sum_{k=1}^{\infty} \beta^{\lceil \frac{k}{B} \rceil - 1} \leq \frac{1}{\beta} \sum_{s=0}^{\infty} \beta^{s} = \frac{B}{\beta(1-\beta)}$.

Proof of Theorem 2.5

If we plug (3.6) into (3.5), we obtain

$$f(\hat{x}_i(k)) \leq f(\hat{y}(k)) + \frac{2mLB}{k\beta(1-\beta)} \sum_{j=1}^{m} \|x_j(0)\| + 2m\alpha L^2 \left(2 + \frac{mB}{\beta(1-\beta)}\right)$$

Finally, the result follows by using the estimate for $f(\hat{y}(k))$ of Lemma (2.4). □