Appendix

A Proof of Theorem 3

Proof. Taking the expectation over the choice of edges (i_k, j_k) gives the following inequality

$$\begin{aligned}
\mathbb{E}_{i_k j_k}[f(x^{k+1})|\eta_k] &\le \mathbb{E}_{i_k j_k}\left[f(x^k) - \frac{1}{4L}\|\nabla_{y_{i_k}} f(x^k) - \nabla_{y_{j_k}} f(x^k)\|^2 - \frac{1}{2L}\|\nabla_{z_{i_k}} f(x^k)\|^2 - \frac{1}{2L}\|\nabla_{z_{j_k}} f(x^k)\|^2\right] \\
&\le f(x^k) - \frac{1}{2}\nabla_y f(x^k)^\top (\mathcal{L}\otimes I_{n_y})\nabla_y f(x^k) - \frac{1}{2}\nabla_z f(x^k)^\top(\mathcal{D}\otimes I_{n_z})\nabla_z f(x^k) \\
&\le f(x^k) - \frac{1}{2}\nabla f(x^k)^\top \mathcal{K}\nabla f(x^k),
\end{aligned} \tag{9}$$

where $\otimes$ denotes the Kronecker product. This shows that the method is a descent method. Now we are ready to prove the main convergence theorem. We have the following:

$$\begin{aligned}
f(x^{k+1}) - f^* &\le \langle \nabla f(x^k), x^k - x^*\rangle \le \|x^k - x^*\|_{\mathcal{K}}^* \|\nabla f(x^k)\|_{\mathcal{K}} \\
&\le R(x^0)\|\nabla f(x^k)\|_{\mathcal{K}} \quad \forall k \ge 0.
\end{aligned}$$

Combining this with inequality (9), we obtain

$$\mathbb{E}[f(x^{k+1}|\eta_k] \le f(x^k) - \frac{(f(x^k) - f^*)^2}{2R^2(x^0)}.$$

Taking the expectation of both sides an denoting $\Delta_k = \mathbb{E}[f(x^k)] - f^*$ gives

$$\Delta_{k+1} \le \Delta_k - \frac{\Delta_k^2}{2R^2(x^0)}.$$

Dividing both sides by $\Delta_k \Delta_{k+1}$ and using the fact that $\Delta_{k+1} \le \Delta_k$ we obtain

$$\frac{1}{\Delta_k} \le \frac{1}{\Delta_{k+1}} - \frac{1}{2R^2(x^0)}.$$

Adding these inequalities for k steps $0 \le \frac{1}{\Delta_0} \le \frac{1}{\Delta_k} - \frac{k}{2R^2(x^0)}$ from which we obtain the statement of the theorem where $C = 2R^2(x^0)$. $\square$

B Proof of Theorem 5

Proof. In this case, the expectation should be over the selection of the pair (i_k, j_k) and random index $l_k \in [N]$. In this proof, the definition of η_k includes l_k i.e., $\eta_k = \{(i_0, j_0, l_0), \dots, (i_{k-1}, j_{k-1}, l_{k-1})\}$. We define the following:

$$\begin{aligned}
d_{i_k}^k &= \left[\frac{\alpha_k}{2L}\left[\nabla_{y_{j_k}} f_{l_k}(x^k) - \nabla_{y_{i_k}} f_{l_k}(x^k)\right]^\top, \quad -\frac{\alpha_k}{L}\left[\nabla_{z_{i_k}} f_{l_k}(x^k)\right]^\top\right]^\top, \\
d_{j_k}^k &= \left[\frac{\alpha_k}{2L}\left[\nabla_{y_{j_k}} f_{l_k}(x^k) - \nabla_{y_{i_k}} f_{l_k}(x^k)\right]^\top, \quad \frac{\alpha_k}{L}\left[\nabla_{z_{j_k}} f_{l_k}(x^k)\right]^\top\right]^\top, \\
d_{i_k j_k}^{l_k} &= U_{i_k} d_{i_k}^k - U_{j_k} d_{j_k}^k.
\end{aligned}$$

For the expectation of objective value at x^{k+1}, we have

$$\begin{aligned}
&\mathbb{E}[f(x^{k+1})|\eta_k] \le \mathbb{E}_{i_k j_k}\mathbb{E}_{l_k}\left[f(x^k) + \left\langle \nabla f(x^k), d_{i_k j_k}^{l_k}\right\rangle + \frac{L}{2}\|d_{i_k j_k}^{l_k}\|^2\right] \\
&\le \mathbb{E}_{i_k j_k}\left[f(x^k) + \left\langle \nabla f(x^k), \mathbb{E}_{l_k}[d_{i_k j_k}^{l_k}]\right\rangle + \frac{L}{2}\mathbb{E}_{l_k}[\|d_{i_k j_k}^{l_k}\|^2]\right] \\
&\le \mathbb{E}_{i_k j_k}\Big[f(x^k) + \frac{\alpha_k}{2L}\left\langle \nabla_{y_{i_k}} f(x^k), \mathbb{E}_{l_k}[\nabla_{y_{j_k}} f_{l_k}(x^k) - \nabla_{y_{i_k}} f_{l_k}(x^k)]\right\rangle \\
&\quad + \frac{\alpha_k}{2L}\left\langle \nabla_{y_{j_k}} f(x^k), \mathbb{E}_{l_k}[\nabla_{y_{i_k}} f_{l_k}(x^k) - \nabla_{y_{j_k}} f_{l_k}(x^k)]\right\rangle \\
&\quad - \frac{\alpha_k}{L}\left\langle \nabla_{z_{i_k}} f(x^k), \mathbb{E}_{l_k}[\nabla_{z_{i_k}} f_{l_k}(x^k)]\right\rangle - \frac{\alpha_k}{L}\left\langle \nabla_{z_{j_k}} f(x^k), \mathbb{E}_{l_k}[\nabla_{z_{j_k}} f_{l_k}(x^k)]\right\rangle + \frac{L}{2}\mathbb{E}_{l_k}[\|d_{i_k j_k}^{l_k}\|^2]\Big].
\end{aligned}$$

Taking expectation over l_k, we get the following relationship:

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})|\eta_k] \leq \mathbb{E}_{i_k j_k}\Big[f(x^k) &+ \frac{\alpha_k}{2L}\left\langle \nabla_{y_{i_k}} f(x^k), \nabla_{y_{j_k}} f(x^k) - \nabla_{y_{i_k}} f(x^k)\right\rangle \\
&+ \frac{\alpha_k}{2L}\left\langle \nabla_{y_{j_k}} f(x^k), \nabla_{y_{i_k}} f(x^k) - \nabla_{y_{j_k}} f(x^k)\right\rangle \\
&- \frac{\alpha_k}{L}\left\langle \nabla_{z_{i_k}} f(x^k), \nabla_{z_{i_k}} f(x^k)\right\rangle - \frac{\alpha_k}{L}\left\langle \nabla_{z_{j_k}} f(x^k), \nabla_{z_{j_k}} f(x^k)\right\rangle + \frac{L}{2}\mathbb{E}_{l_k}[\|d^{l_k}_{i_k j_k}\|^2]\Big].
\end{aligned}$$

We first note that $\mathbb{E}_{l_k}[\|d^{l_k}_{i_k j_k}\|^2] \leq 8M^2\alpha_k^2/L^2$ since $\|\nabla f_l\| \leq M$. Substituting this in the above inequality and simplifying we get,

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})|\eta_k] &\leq f(x^k) - \alpha_k \nabla_y f(x^k)^\top(\mathcal{L}\otimes I_n)\nabla_y f(x^k) - \alpha_k \nabla_z f(x^k)^\top(\mathcal{D}\otimes I_n)\nabla_z f(x^k) + \frac{4M^2\alpha_k^2}{L} \\
&\leq f(x^k) - \alpha_k \nabla f(x^k)^\top \mathcal{K}\nabla f(x^k) + \frac{4M^2\alpha_k^2}{L}. \qquad (10)
\end{aligned}$$

Similar to Theorem 3, we obtain a lower bound on $\nabla f(x^k)^\top\mathcal{K}\nabla f(x^k)$ in the following manner.

$$\begin{aligned}
f(x^k) - f^* &\leq \langle \nabla f(x^k), x^k - x^*\rangle \leq \|x^k - x^*\|^*_{\mathcal{K}}.\|\nabla f(x^k)\|_{\mathcal{K}} \\
&\leq R(x^0)\|\nabla f(x^k)\|_{\mathcal{K}}.
\end{aligned}$$

Combining this with inequality Equation 10, we obtain

$$\mathbb{E}[f(x_{k+1})|\eta_k] \leq f(x^k) - \alpha_k\frac{(f(x^k)-f^*)^2}{R^2(x^0)} + \frac{4M^2\alpha_k^2}{L}.$$

Taking the expectation of both sides an denoting $\Delta_k = \mathbb{E}[f(x^k)] - f^*$ gives

$$\Delta_{k+1} \leq \Delta_k - \alpha_k\frac{\Delta_k^2}{R^2(x^0)} + \frac{4M^2\alpha_k^2}{L}.$$

Adding these inequalities from $i=0$ to $i=k$ and use telescopy we get,

$$\Delta_{k+1} + \sum_{i=0}^{k}\alpha_i\frac{\Delta_k^2}{R^2(x^0)} \leq \Delta_0 + \frac{4M^2}{L}\sum_{i=0}^{k}\alpha_i^2.$$

Using the definition of $\bar{x}_{k+1} = \arg\min_{0\leq i\leq k+1} f(x_i)$, we get

$$\sum_{i=0}^{k}\alpha_i\frac{(\mathbb{E}[f(\bar{x}_{k+1})-f^*])^2}{R^2(x^0)} \leq \Delta_{k+1} + \sum_{i=0}^{k}\alpha_i\frac{\Delta_k^2}{R^2(x^0)} \leq \Delta_0 + \frac{4M^2}{L}\sum_{i=0}^{k}\alpha_i^2.$$

Therefore, from the above inequality we have,

$$\mathbb{E}[f(\bar{x}_{k+1})-f^*] \leq R(x^0)\sqrt{\frac{(\Delta_0 + 4M^2\sum_{i=0}^{k}\alpha_i^2/L)}{\sum_{i=0}^{k}\alpha_i}}.$$

Note that $\mathbb{E}[f(\bar{x}_{k+1})-f^*] \to 0$ if we choose step sizes satisfying the condition that $\sum_{i=0}^{\infty}\alpha_i = \infty$ and $\sum_{i=0}^{\infty}\alpha_i^2 < \infty$. Substituting $\alpha_i = \sqrt{\Delta_0 L}/(2M\sqrt{i+1})$, we get the required result using the reasoning from [24] (we refer the reader to Section 2.2 of [24] for more details). ☐

C Proof of Theorem 4

Proof. For ease of exposition, we analyze the case where the unconstrained variables z are absent. The analysis of case with z variables can be carried out in a similar manner. Consider the update on edge (i_k, j_k). Recall that $D(k)$ denotes the index of the iterate used in the k^{th} iteration for calculating the gradients. Let $d^k = \frac{\alpha_k}{2L}\left(\nabla_{y_{j_k}} f(x^{D(k)}) - \nabla_{y_{i_k}} f(x^{D(k)})\right)$

and $d^k_{i_k j_k} = x^{k+1} - x^k = U_{i_k} d^k - U_{j_k} d^k$. Note that $\|d^k_{i_k j_k}\|^2 = 2\|d^k\|^2$. Since f is Lipschitz continuous gradient, we have

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \left\langle \nabla_{y_{i_k} y_{j_k}} f(x^k), d^k_{i_k j_k} \right\rangle + \frac{L}{2}\|d^k_{i_k j_k}\|^2 \\
&\leq f(x^k) + \left\langle \nabla_{y_{i_k} y_{j_k}} f(x^{D(k)}) + \nabla_{y_{i_k} y_{j_k}} f(x^k) - \nabla_{y_{i_k} y_{j_k}} f(x^{D(k)}), d^k_{i_k j_k} \right\rangle + \frac{L}{2}\|d^k_{i_k j_k}\|^2 \\
&\leq f(x^k) - \frac{L}{\alpha_k}\|d^k_{i_k j_k}\|^2 + \left\langle \nabla_{y_{i_k} y_{j_k}} f(x^k) - \nabla_{y_{i_k} y_{j_k}} f(x^{D(k)}), d^k_{i_k j_k} \right\rangle + \frac{L}{2}\|d^k_{i_k j_k}\|^2 \\
&\leq f(x^k) - L\left(\frac{1}{\alpha_k} - \frac{1}{2}\right)\|d^k_{i_k j_k}\|^2 + \|\nabla_{y_{i_k} y_{j_k}} f(x^k) - \nabla_{y_{i_k} y_{j_k}} f(x^{D(k)})\|\|d^k_{i_k j_k}\| \\
&\leq f(x^k) - L\left(\frac{1}{\alpha_k} - \frac{1}{2}\right)\|d^k_{i_k j_k}\|^2 + L\|x^k - x^{D(k)}\|\|d^k_{i_k j_k}\|.
\end{aligned}$$

The third and fourth steps in the above derivation follow from definition of d^k_{ij} and Cauchy-Schwarz inequality respectively. The last step follows from the fact the gradients are Lipschitz continuous. Using the assumption that staleness in the variables is bounded by τ, i.e., $k - D(k) \leq \tau$ and definition of d^k_{ij}, we have

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - L\left(\frac{1}{\alpha_k} - \frac{1}{2}\right)\|d^k_{i_k j_k}\|^2 + L\left(\sum_{t=1}^{\tau} \|d^{k-t}_{i_{k-t} j_{k-t}}\|\|d^k_{i_k j_k}\|\right) \\
&\leq f(x^k) - L\left(\frac{1}{\alpha_k} - \frac{1}{2}\right)\|d^k_{i_k j_k}\|^2 + \frac{L}{2}\left(\sum_{t=1}^{\tau} \left[\|d^{k-t}_{i_{k-t} j_{k-t}}\|^2 + \|d^k_{i_k j_k}\|^2\right]\right) \\
&\leq f(x^k) - L\left(\frac{1}{\alpha_k} - \frac{1+\tau}{2}\right)\|d^k_{i_k j_k}\|^2 + \frac{L}{2}\sum_{t=1}^{\tau} \|d^{k-t}_{i_{k-t} j_{k-t}}\|^2.
\end{aligned}$$

The first step follows from triangle inequality. The second inequality follows from fact that $ab \leq (a^2 + b^2)/2$. Using expectation over the edges, we have

$$\mathbb{E}[f(x^{k+1})] \leq \mathbb{E}[f(x^k)] - L\left(\frac{1}{\alpha_k} - \frac{1+\tau}{2}\right)\mathbb{E}[\|d^k_{i_k j_k}\|^2] + \frac{L}{2}\mathbb{E}\left[\sum_{t=1}^{\tau} \|d^{k-t}_{i_{k-t} j_{k-t}}\|^2\right]. \tag{11}$$

We now prove that, for all $k \geq 0$

$$\mathbb{E}\left[\|d^{k-1}_{i_{k-1} j_{k-1}}\|^2\right] \leq \rho\mathbb{E}\left[\|d^k_{i_k j_k}\|^2\right], \tag{12}$$

where we define $\mathbb{E}\left[\|d^{k-1}_{i_{k-1} j_{k-1}}\|^2\right] = 0$ for $k = 0$. Let w^t denote the vector of size $|E|$ such that $w^t_{ij} = \sqrt{p_{ij}}\|d^t_{ij}\|$ (with slight abuse of notation, we use w^t_{ij} to denote the entry corresponding to edge (i, j)). Note that $\mathbb{E}\left[\|d^t_{i_t j_t}\|^2\right] = \mathbb{E}[\|w^t\|^2]$. We prove Equation (12) by induction.

Let u^k be a vector of size $|E|$ such that $u^k_{ij} = \sqrt{p_{ij}}\|d^k_{ij} - d^{k-1}_{ij}\|$. Consider the following:

$$\begin{aligned}
\mathbb{E}[\|w^{k-1}\|]^2 - \mathbb{E}[\|w^k\|^2] &= \mathbb{E}[2\|w^{k-1}\|]^2 - \mathbb{E}[\|w^k\|^2 + \|w^{k-1}\|^2] \\
&\leq 2\mathbb{E}[\|w^{k-1}\|^2] - 2\mathbb{E}[\langle w^{k-1}, w^k\rangle] \\
&\leq 2\mathbb{E}[\|w^{k-1}\|\|w^{k-1} - w^k\|] \\
&\leq 2\mathbb{E}[\|w^{k-1}\|\|u^k\|] \leq 2\mathbb{E}[\|w^{k-1}\|\sqrt{2}\alpha_k\|x^{D(k)} - x^{D(k-1)}\|] \\
&\leq \sqrt{2}\alpha_k \sum_{t=\min(D(k-1),D(k))}^{\max(D(k-1),D(k))} \left(\mathbb{E}[\|w^{k-1}\|^2] + \mathbb{E}[\|d^t_{i_t j_t}\|^2]\right).
\end{aligned} \tag{13}$$

The fourth step follows from the bound below on $|u_{ij}^k|$

$$\begin{aligned}
|u_{ij}^k| &= \sqrt{p_{ij}}\|d_{ij}^k - d_{ij}^{k-1}\| \\
&\leq \sqrt{p_{ij}}\|(U_i - U_j)\frac{\alpha_k}{2L}(\nabla_{y_i} f(x^{D(k)}) - \nabla_{y_j} f(x^{D(k)}) + \nabla_{y_j} f(x^{D(k-1)}) - \nabla_{y_i} f(x^{D(k-1)}))\| \\
&\leq \sqrt{2p_{ij}}\alpha_k\|x^{D(k)} - x^{D(k-1)}\|.
\end{aligned}$$

The fifth step follows from triangle inequality. We now prove (12): the induction hypothesis is trivially true for $k = 0$. Assume it is true for some $k - 1 \geq 0$. Now using Equation (13), we have

$$\mathbb{E}[\|w^{k-1}\|]^2 - \mathbb{E}[\|w^k\|^2] \leq \sqrt{2}\alpha_k(\tau+2)\mathbb{E}[\|w^{k-1}\|^2] + \sqrt{2}\alpha_k(\tau+2)\rho^{\tau+1}\mathbb{E}[\|w^k\|^2]$$

for our choice of α_k. The last step follows from the fact that $\mathbb{E}[\|d_{i_t j_t}^t\|^2] = \mathbb{E}[\|w^t\|^2]$ and mathematical induction. From the above, we get

$$\mathbb{E}[\|w^{k-1}\|^2] \leq \frac{1 + \sqrt{2}\alpha_k(\tau+2)\rho^{(\tau+1)}}{1 - \sqrt{2}\alpha_k(\tau+2)}\mathbb{E}[\|w^k\|^2] \leq \rho\mathbb{E}[\|w^k\|^2].$$

Thus, the statement holds for k. Therefore, the statement holds for all $k \in \mathbb{N}$ by mathematical induction. Substituting the above in Equation (11), we get

$$\mathbb{E}[f(x^{k+1})] \leq \mathbb{E}[f(x^k)] - L\left(\frac{1}{\alpha_k} - \frac{1+\tau+\tau\rho^\tau}{2}\right)\mathbb{E}[\|d_{i_k j_k}^k\|^2].$$

This proves that the method is a descent method in expectation. Using the definition of d_{ij}^k, we have

$$\begin{aligned}
\mathbb{E}[f(x^{k+1})] &\leq \mathbb{E}[f(x^k)] - \frac{\alpha_k^2}{4L}\left(\frac{1}{\alpha_k} - \frac{1+\tau+\tau\rho^\tau}{2}\right)\mathbb{E}[\|\nabla_{y_{i_k}} f(x^{D(k)}) - \nabla_{y_{j_k}} f(x^{D(k)})\|^2] \\
&\leq \mathbb{E}[f(x^k)] - \frac{\alpha_k^2}{4L}\left(\frac{1}{\alpha_k} - \frac{1+\tau+\tau\rho^\tau}{2}\right)\mathbb{E}[\|\nabla f(x^{D(k)}) - \nabla f(x^{D(k)})\|_{\mathcal{K}}^2] \\
&\leq \mathbb{E}[f(x^k)] - \frac{\alpha_k^2}{2R^2(x^0)}\left(\frac{1}{\alpha_k} - \frac{1+\tau+\tau\rho^\tau}{2}\right)\mathbb{E}[(f(x^{D(k)}) - f^*)^2] \\
&\leq \mathbb{E}[f(x^k)] - \frac{\alpha_k^2}{2R^2(x^0)}\left(\frac{1}{\alpha_k} - \frac{1+\tau+\tau\rho^\tau}{2}\right)\mathbb{E}[(f(x^k) - f^*)^2].
\end{aligned}$$

The second and third steps are similar to the proof of Theorem 3. The last step follows from the fact that the method is a descent method in expectation. Following similar analysis as Theorem 3, we get the required result. □

D Proof of Theorem 6

Proof. Let $Ax = \sum_i x_i$. Let $\tilde{x}_{k+1}$ be solution to the following optimization problem:

$$\tilde{x}^{k+1} = \arg\min_{\{x|Ax=0\}} \langle\nabla f(x^k), x - x^k\rangle + \frac{L}{2}\|x - x^k\|^2 + h(x).$$

To prove our result, we first prove few intermediate results. We say vectors $d \in \mathbb{R}^n$ and $d' \in \mathbb{R}^n$ are conformal if $d_i d_i' \geq 0$ for all $i \in [b]$. We use $d_{i_k j_k} = x^{k+1} - x^k$ and $d = \tilde{x}^{k+1} - x^k$. Our first claim is that for any d, we can always find conformal vectors whose sum is d (see [22]). More formally, we have the following result.

Lemma 7. *For any $d \in \mathbb{R}^n$ with $Ad = 0$, we have a multi-set $S = \{d_{ij}'\}_{i\neq j}$ such that d and d_{ij}' are conformal for all $i \neq j$ and $i, j \in [b]$ i.e., $\sum_{i\neq j} d_{ij}' = d$, $Ad_{ij}' = 0$ and d_{ij}' can be non-zero only in coordinates corresponding to x_i and x_j.*

Proof. We prove by an iterative construction, i.e., for every vector d such that $Ad = 0$, we construct a set $S = \{s_{ij}\}$ ($s_{ij} \in \mathbb{R}^n$) with the required properties. We start with a vector $u^0 = d$ and multi-set $S^0 = \{s_{ij}^0\}$ and $s_{ij}^0 = 0$ for all $i \neq j$ and $i, j \in [n]$. At the k^{th} step of the construction, we will have $Au^k = 0$, $As = 0$ for all $s \in S^k$, $d = u^k + \sum_{s\in S^k} s$ and each element of s is conformal to d.

In k^{th} iteration, pick the element with the smallest absolute value (say v) in u^{k-1}. Let us assume it corresponds to y_p^j. Now pick an element from u^{k-1} corresponding to y_q^j for $p \neq q \in [m]$ with at least absolute value v albeit with opposite sign. Note that such an element should exist since $Au^{k-1} = 0$. Let p_1 and p_2 denote the indices of these elements in u^{k-1}. Let S^k be same as S^{k-1} except for s_{pq}^k which is given by $s_{pq}^k = s_{pq}^{k-1} + r = s_{pq}^{k-1} + u_{p_1}^{k-1} e_{p_1} - u_{p_1}^{k-1} e_{p_2}$ where e_i denotes a vector in $\mathbb{R}^n$ with zero in all components except in i^{th} position (where it is one). Note that $Ar = 0$ and r is conformal to d since it has the same sign. Let $u^{k+1} = u^k - r$. Note that $Au^{k+1} = 0$ since $Au^k = 0$ and $Ar = 0$. Also observe that $As = 0$ for all $s \in S^{k+1}$ and $u^{k+1} = \sum_{s \in S^k} s = d$.

Finally, note that each iteration the number of non-zero elements of u^k decrease by at least 1. Therefore, this algorithm terminates after a finite number of iterations. Moreover, at termination $u^k = 0$ otherwise the algorithm can always pick an element and continue with the process. This gives us the required conformal multi-set. □

Now consider a set $\{d'_{ij}\}$ which is conformal to d. We define $\hat{x}_{k+1}$ in the following manner:

$$\hat{x}_i^{k+1} = \begin{cases} x_i^k + d'_{ij} & \text{if } (i,j) = (i_k, j_k) \\ x_i^k & \text{if } (i,j) \neq (i_k, j_k) \end{cases}$$

Lemma 8. *For any $x \in \mathbb{R}^n$ and $k \geq 0$,*

$$\mathbb{E}[\|\hat{x}^{k+1} - x^k\|^2 \leq \lambda(\|\tilde{x}^{k+1} - x^k\|^2).$$

We also have

$$\mathbb{E}(h(\hat{x}^{k+1})) \leq (1-\lambda)h(x^k) + \lambda h(\tilde{x}^{k+1}).$$

Proof. We have the following bound:

$$\mathbb{E}_{i_k j_k}[\|\hat{x}^{k+1} - x^k\|^2 = \lambda \sum_{i \neq j} \|d'_{ij}\|^2 \leq \lambda \| \sum_{i \neq j} d'_{ij}\|^2 = \lambda \|d\|^2 = \lambda \|\tilde{x}^{k+1} - x^k\|^2.$$

The above statement directly follows the fact that $\{d'_{ij}\}$ is conformal to d. The remaining part directly follows from [22]. □

The remaining part essentially on similar lines as [22]. We give the details here for completeness. From Lemma 1, we have

$$\begin{aligned}
\mathbb{E}_{i_k j_k}[F(x^{k+1})] &\leq \mathbb{E}_{i_k j_k}[f(x^k) + \langle \nabla f(x^k), d_{i_k j_k} \rangle + \frac{L}{2}\|d_{i_k j_k}\|^2 + h(x^k + d_{i_k j_k})] \\
&\leq \mathbb{E}_{i_k j_k}[f(x^k) + \langle \nabla f(x^k), d'_{i_k j_k} \rangle + \frac{L}{2}\|d'_{i_k j_k}\|^2 + h(x^k + d'_{i_k j_k})] \\
&= f(x^k) + \lambda \left(\langle \nabla f(x), \sum_{i \neq j} d'_{ij} \rangle + \sum_{i \neq j} \frac{L}{2}\|d'_{ij}\|^2 + \sum_{i \neq j} h(x + d'_{ij}) \right) \\
&\leq (1-\lambda)F(x^k) + \lambda(f(x^k) + \langle \nabla f(x), d \rangle + \frac{L}{2}\|d\|^2 + h(x + d)) \\
&\leq \min_{\{y | Ay = 0\}} (1-\lambda)F(x^k) + \lambda(F(y) + \frac{L}{2}\|y - x^k\|^2) \\
&\leq \min_{\beta \in [0,1]} (1-\lambda)F(x^k) + \lambda(F(\beta x^* + (1-\beta)x^k) + \frac{\beta^2 L}{2}\|x^k - x^*\|^2) \\
&\leq (1-\lambda)F(x^k) + \lambda \left(F(x^k) - \frac{2(F(x^k) - F(x^*))^2}{LR^2(x^0)} \right).
\end{aligned}$$

The second step follows from optimality of $d_{i_k j_k}$. The fourth step follows from Lemma 8. Now using the similar recurrence relation as in Theorem 2, we get the required result. □

E Reduction of General Case

In this section we show how to reduce a problem with linear constraints to the form of Problem 4 in the paper. For simplicity, we focus on smooth objective functions. However, the formulation can be extended to composite objective functions along similar lines. Consider the optimization problem

$$\min_{x} f(x)$$
$$s.t.\ Ax = \sum A_i x_i = 0,$$

where f_i is a convex function with an L-Lipschitz gradient.

Let $\bar{A}_i$ be a matrix with orthonormal columns satisfying $\mathrm{range}(\bar{A}_i) = \ker(A_i)$, this can be obtained (e.g. using SVD). For each i, define $y_i = A_i x_i$ and assume that the rank of A_i is less than or equal to the dimensionality of x_i. [4] Then we can rewrite x as a function $h(y, z)$ satisfying

$$x_i = A_i^+ y_i + \bar{A}_i z_i,$$

for some unknown z_i, where C^+ denote the pseudo-inverse of C. The problem then becomes

$$\min_{y,z} g(y, z) \ \ s.t. \ \sum_{i=1}^{N} y_i = 0, \tag{14}$$

where

$$g(y, z) = f(\phi(y, z)) = f\left(\sum_i U_i(A_i^+ y_i + \bar{A}_i z_i)\right). \tag{15}$$

It is clear that the sets $S_1 = \{x|Ax = 0\}$ and $S_2 = \{\phi(y, z)| \sum_i y_i = 0\}$ are equal and hence the problem defined in 14 is equivalent to that in 1.

Note that such a transformation preserves convexity of the objective function. It is also easy to show that it preserves the block-wise Lipschitz continuity of the gradients as we prove in the following result.

Lemma 9. *Let f be a function with L_i-Lipschitz gradient w.r.t x_i. Let $g(y, z)$ be the function defined in* 15*. Then g satisfies the following condition*

$$\|\nabla_{y_i} g(y, z) - \nabla_{y_i} g(y', z)\| \leq \frac{L_i}{\sigma^2_{\min}(A_i)} \|y_i - y'_i\|$$
$$\|\nabla_{z_i} g(y, z) - \nabla_{z_i} g(y, z')\| \leq L_i \|z_i - z'_i\|,$$

where $\sigma_{\min}(B)$ denotes the minimum non-zero singular value of B.

Proof. We have

$$\begin{aligned}\|\nabla_{y_i} g(y, z) - \nabla_{y_i} g(y', z)\| &= \|(U_i A_i^+)^\top [\nabla_x f(\phi(y, z)) - \nabla_x f(\phi(y', z))]\| \\ &\leq \|A_i^+\| \|\nabla_i f(\phi(y, z)) - \nabla_i f(\phi(y', z))\| \\ &\leq L_i \|A_i^+\| \|A_i^+(y_i - y'_i)\| \leq L_i \|A_i^+\|^2 \|y_i - y'_i\| = \frac{L_i}{\sigma^2_{\min}(A_i)} \|y_i - y'_i\|,\end{aligned}$$

Similar proof holds for $\|\nabla_{z_i} g(y, z) - \nabla_{z_i} g(y, z')\|$, noting that $\|\bar{A}_i\| = 1$. □

It is worth noting that this reduction is mainly used to simplify analysis. In practice, however, we observed that an algorithm that operates directly on the original variables x_i (i.e. Algorithm 1) converges much faster and is much less sensitive to the conditioning of A_i compared to an algorithm that operates on y_i and z_i. Indeed, with appropriate step sizes, Algorithm 1 minimizes, in each step, a tighter bound on the objective function compared to the bound based 14 as stated in the following result.

[4] If the rank constraint is not satisfied then one solution is to use a coarser partitioning of x so that the dimensionality of x_i is large enough.

Lemma 10. *Let g and ϕ be as defined in 15. And let*

$$d_i = A_i^{+} d_{y_i} + \bar{A}_i d_{z_i}.$$

Then, for any d_i and d_j satisfying $A_i d_i + A_j d_j = 0$ and any feasible $x = \phi(y, z)$ we have

$$\begin{aligned}
&\langle \nabla_i f(x), d_i \rangle + \langle \nabla_j f(x), d_j \rangle + \frac{L_i}{2\alpha}\|d_i\|^2 + \frac{L_j}{2\alpha}\|d_j\|^2 \\
&\leq \langle \nabla_{y_i} g(y, z), d_{y_i} \rangle + \langle \nabla_{z_i} g(y, z), d_{z_i} \rangle + \langle \nabla_{y_j} g(y, z), d_{y_j} \rangle + \langle \nabla_{z_j} g(y, z), d_{z_j} \rangle \\
&+ \frac{L_i}{2\alpha\sigma_{\min}^2(A_i)}\|d_{y_i}\|^2 + \frac{L_i}{2\alpha}\|d_{z_i}\|^2 + \frac{L_j}{2\alpha\sigma_{\min}^2(A_j)}\|d_{y_j}\|^2 + \frac{L_j}{2\alpha}\|d_{z_j}\|^2.
\end{aligned}$$

Proof. The proof follows directly from the fact that

$$\nabla_i f(x) = {A_i^{+}}^{\top} \nabla_{y_i} g(y, z) + {\bar{A}_i}^{\top} \nabla_{z_i} g(y, z).$$

□