A Basic Inequalities

Firstly, we present some preliminary inequalities that will be frequently used in the subsequent proofs.

Fact 1 (Cauchy-Schwarz Inequality). For any \( x, y \in \mathbb{R}^d \),
\[
|\langle x, y \rangle| \leq \| x \| \cdot \| y \|. \tag{15}
\]

Fact 2. For arbitrary set of \( N \) vectors \( \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d \),
\[
\left\| \sum_{i=1}^N x_i \right\|^2 \leq N \sum_{i=1}^N \| x_i \|^2. \tag{16}
\]

Fact 3. For any \( x, y \in \mathbb{R}^d \),
\[
\| x + y \|^2 \leq (1 + \alpha) \| x \|^2 + (1 + \alpha^{-1}) \| y \|^2, \quad \forall \alpha > 0. \tag{17}
\]

Fact 4. Given a convex set \( \mathcal{K} \in \mathbb{R}^d \), the projection operator satisfies the following properties
\begin{align*}
\text{(i)} & \quad \| P_\mathcal{K}(x) - P_\mathcal{K}(y) \| \leq \| x - y \|, \quad \forall x, y \in \mathbb{R}^d. \tag{18} \\
\text{(ii)} & \quad \| P_\mathcal{K}(x) - x \| \leq \| x - y \|, \quad \forall x \in \mathbb{R}^d, y \in \mathcal{K}. \tag{19} \\
\text{(iii)} & \quad (P_\mathcal{K}(x) - x, x - y) \leq -\| P_\mathcal{K}(x) - x \|^2 \leq 0, \quad \forall x \in \mathbb{R}^d, y \in \mathcal{K}. \tag{20}
\end{align*}

Fact 5 (Jensen’s Inequality). Given a convex function \( f \) and a random variable \( x \), then
\[
f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]. \tag{21}
\]

B Proofs of Section 2

Define
\[
\hat{x}_i^{t+1} := x_i^t + \gamma \sum_{j \in N_i} a_{ij} (\hat{x}_j^t - \hat{x}_i^t) - \eta_i \nabla f_i^t(x_i^t),
\]
\[
r_i^{t+1} := P_\mathcal{K} (\hat{x}_i^{t+1}) - \hat{x}_i^{t+1},
\]
\[
\bar{x}^t := \frac{1}{N} \sum_{i=1}^N x_i^t,
\]
and then
\[
x_i^{t+1} = P_\mathcal{K} (\hat{x}_i^{t+1}) = \bar{x}^{t+1} + r_i^{t+1}. \tag{25}
\]

For notational simplicity, define matrices
\[
X^t := \text{col}\{x_1^t, \ldots, x_N^t\}, \quad \bar{X}^t := \text{col}\{\bar{x}_1^t, \ldots, \bar{x}_N^t\}, \quad \hat{X}_i^t := \text{col}\{\hat{x}_1^t, \ldots, \hat{x}_N^t\},
\]
\[
R^t := \text{col}\{r_1^t, \ldots, r_N^t\}, \quad \nabla F^t(X^t) := \text{col}\{\nabla f_1^t(x_1^t), \ldots, \nabla f_N^t(x_N^t)\}.
\]

Denote by \( 1_N \) the \( N \)-dimensional column vector with all components being one, and \( M := \frac{1}{N} 1_N 1_N^T \), \( M := M \otimes I_d \). Then \( \bar{X}^t = M X^t \). Define the Laplacian matrix \( L := I_N - A \) and \( L := L \otimes I_d, I := I_N \otimes I_d \), where \( \otimes \) is the Kronecker product. Denote by \( L_i \) the \( i \)-th row of \( L \). Then by Remark 1, Algorithm 1 can be written in the matrix form as
\begin{align*}
\dot{X}^{t+1} &= \dot{X}^t + Q(X^t - \bar{X}^t), \tag{26} \\
X^{t+1} &= P_\mathcal{K} \left( X^t - \eta_t \dot{X}^{t+1} - \nabla F^t(X^t) \right) \tag{27} \\
&= \bar{X}^{t+1} + R^{t+1}. \tag{28}
\end{align*}

To begin with, we consider general regret bounds.
Lemma 1. Consider Algorithm $\text{[Algorithm]}$ with non-increasing gradient descent stepsizes $\{\eta_t\}_{t=1}^{T}$.

(i) (Convex case) Suppose Assumptions $1, 3, 4$ hold. Then for each $j \in V$:

$$R(j, T) \leq \frac{ND^2}{2\eta_T} + NG^2 \sum_{t=1}^{T} \eta_t + (2\sqrt{N} + N) G \sum_{t=1}^{T} \|X^t - \tilde{X}^t\|$$

$$+ \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\tilde{X}^t - \tilde{X}^{t+1}\|^2 + 3\|R^{t+1}\|^2\right). \tag{29}$$

(ii) (Strongly convex case) Suppose Assumptions $1, 2, 4$ hold and $\eta_t = \frac{1}{\mu(t+c)}$ for a constant $c \geq 0$. Then for each $j \in V$:

$$R(j, T) \leq \mu c D^2 + NG^2 \sum_{t=1}^{T} \eta_t + (2\sqrt{N} + N) G \sum_{t=1}^{T} \|X^t - \tilde{X}^t\|$$

$$+ \sum_{t=1}^{T} \frac{1}{2\eta_t} \left(\|\tilde{X}^t - \tilde{X}^{t+1}\|^2 + 3\|R^{t+1}\|^2\right). \tag{30}$$

Proof. Because $\sum_{i=1}^{N} \sum_{j \in V_i} a_{ij} (\tilde{x}^{t+1}_j - \tilde{x}^{t+1}_i) = 0$ under Assumption $1$ with the introduction of the projection error $r^{t+1}_i$, we can write

$$\tilde{x}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} (\tilde{x}^{t+1}_i + r^{t+1}_i) = \tilde{x}^t - \frac{\eta_t}{N} \sum_{i=1}^{N} \nabla f^t_i(x^t_i) + \frac{1}{N} \sum_{i=1}^{N} r^{t+1}_i.$$ Denote by $x^*$ the best decision in the hindsight, i.e., $x^* = \arg \min_{x \in \mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} f^t_i(x)$.

Then, \begin{equation} \|\tilde{x}^{t+1} - x^*\|^2 = \|x^t - x^*\|^2 + \frac{1}{N^2} \left(\sum_{i=1}^{N} r^{t+1}_i - \eta_t \sum_{i=1}^{N} \nabla f^t_i(x^t_i)\right)^2 + \frac{2}{N} \sum_{i=1}^{N} \langle r^{t+1}_i, x^t - x^* \rangle - \frac{2\eta_t}{N} \sum_{i=1}^{N} \langle \nabla f^t_i(x^t_i), \tilde{x}^t - x^* \rangle. \tag{31} \end{equation}

Under Assumption $4$ we estimate the second term

$$\frac{1}{N^2} \left(\sum_{i=1}^{N} r^{t+1}_i - \eta_t \sum_{i=1}^{N} \nabla f^t_i(x^t_i)\right)^2 = \frac{1}{N^2} \left(2 \left\|\sum_{i=1}^{N} r^{t+1}_i\right\|^2 + 2\eta_t^2 \left\|\sum_{i=1}^{N} \nabla f^t_i(x^t_i)\right\|^2\right) \tag{30}$$

$$\leq \frac{1}{N^2} \left(2N \sum_{i=1}^{N} \|r^{t+1}_i\|^2 + 2\eta_t^2 N^2 G^2\right) = \frac{2}{N} \|R^{t+1}\|^2 + 2\eta_t^2 G^2. \tag{32}$$

Then we come to the third term. Noting that $x^* \in \mathcal{K}$, by using the definition of $r^{t+1}_i$ and the projection property (iii), we have

$$\sum_{i=1}^{N} \langle r^{t+1}_i, x^t - x^* \rangle = \sum_{i=1}^{N} \left(\langle r^{t+1}_i, \tilde{x}^t - \tilde{x}^{t+1}_i \rangle + \langle P_{\mathcal{K}} (\tilde{x}^{t+1}_i) - \tilde{x}^{t+1}_i, x^t - x^* \rangle\right) \tag{33}$$

$$\leq \sum_{i=1}^{N} \|r^{t+1}_i\| + \|\tilde{x}^t - \tilde{x}^{t+1}\|^2 \leq \frac{1}{2} \left(\|r^{t+1}_i\|^2 + \|\tilde{x}^t - \tilde{x}^{t+1}\|^2\right) = \frac{1}{2} \|R^{t+1}\|^2 + \|\tilde{X}^t - \tilde{X}^{t+1}\|^2. \tag{30}$$

Next we turn to the fourth term. Under Assumption $4$,

$$f^t_i(x^t_i) \geq f^t_i(x^t_i), \quad \langle \nabla f^t_i(x^t_i), x^t - x^{t+1}_i \rangle \geq f^t_i(x^{t+1}_i) - G \|x^t_i - x^{t+1}_i\|,$$

and hence,

$$-\langle \nabla f^t_i(x^t_i), x^t - x^* \rangle = \langle \nabla f^t_i(x^t_i), x^* - x^t_i \rangle + \langle \nabla f^t_i(x^t_i), x^t_i - x^t \rangle$$

$$\leq f^t_i(x^t) - f^t_i(x^t_i) - \frac{\mu}{2} \|x^* - x^t_i\|^2 + G \|x^t_i - x^t\|$$

$$\leq f^t_i(x^*) - f^t_i(x^{t+1}_i) + G \|x^t_i - x^*\| - \frac{\mu}{2} \|x^t_i - x^{t+1}_i\|^2 + G \|x^t_i - x^t\|. \tag{34}$$
where $\mu > 0$ for the strongly convex case and $\mu \equiv 0$ for the convex case. Summing up (34) over $i = 1, \cdots , N$ with the fact that
\[
\sum_{i=1}^N \|x_i^t - x_j^t\| \leq \sum_{i=1}^N \|x_i^t - \bar{x}\| + N\|\bar{x} - x_j^t\| \leq \sqrt{N}\|X^t - \bar{X}\| + N\|X^t - \bar{X}\|, \\
\sum_{i=1}^N \|x_i^t - x_j^t\|^2 \geq \frac{1}{N} \left( \sum_{i=1}^N (x_i^t - x_i^t) \right)^2 \geq \frac{1}{N} \|N\bar{x}^t - N\bar{x}\|^2 = N\|x^* - \bar{x}\|^2, 
\]
we have
\[
- \sum_{i=1}^N \langle \nabla f_i^t(x_i^t), \bar{x}^t - x^* \rangle \leq \sum_{i=1}^N (f_i^t(x_i^t) - f_i^t(x_j^t)) + (2\sqrt{N+N})G\|X^t - \bar{X}\| - \frac{N\mu}{2} \|x^* - \bar{x}\|^2. 
\]
By substituting (32), (33), and (35) into (31), we derive
\[
\|\bar{x}^{t+1} - x^*\|^2 \leq \|\bar{x}^t - x^*\|^2 + \frac{2}{N} \|R^{t+1}\|^2 + 2\eta_t^2G^2 + \frac{1}{N} \left( \|R^{t+1}\|^2 \|X^t - \bar{X}^{t+1}\|^2 \right) \\
+ \frac{2\eta_t}{N} \left( \sum_{i=1}^N (f_i^t(x_i^t) - f_i^t(x_j^t)) + (2\sqrt{N+N})G\|X^t - \bar{X}^t\| - \frac{N\mu}{2} \|x^* - \bar{x}\|^2 \right).
\]
By rearranging the terms,
\[
\sum_{i=1}^N (f_i^t(x_i^t) - f_i^t(x^*)) \leq \frac{N}{2} \left( \frac{1}{\eta_t} - \mu \right) \|\bar{x}^t - x^*\|^2 - \frac{1}{\eta_t} \|\bar{x}^{t+1} - x^*\|^2 \\
+ NG^2\eta_t + (2\sqrt{N+N})G\|X^t - \bar{X}^t\| + \frac{1}{2\eta_t} \left( \|\bar{x}^t - \bar{x}^{t+1}\|^2 + 3\|R^{t+1}\|^2 \right).
\]
Summing up the above inequality over $t = 1, \cdots , T$ gives
\[
\sum_{t=1}^T \sum_{i=1}^N (f_i^t(x_i^t) - f_i^t(x^*)) \leq \frac{N}{2} \sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \mu \right) \|\bar{x}^t - x^*\|^2 + NG^2 \sum_{t=1}^T \eta_t \\
+ (2\sqrt{N+N})G \sum_{t=1}^T \|X^t - \bar{X}^t\| + \sum_{t=1}^T \frac{1}{2\eta_t} \left( \|\bar{x}^t - \bar{x}^{t+1}\|^2 + 3\|R^{t+1}\|^2 \right), \frac{1}{\eta_0} \equiv 0 
\]
(i) In the convex case, $\mu \equiv 0$. Using Assumption 5 with the non-increasing of $\{\eta_t\}_{t=1}^T$, we have
\[
\sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \|\bar{x}^t - x^*\|^2 \leq \sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) D^2 \frac{D^2}{\eta_T}. 
\]
By substituting (37) into (36), we derive (29).
(ii) Under Assumption 5, $\mu > 0$, and thus, $\eta_t = \frac{1}{\mu(t+\tau)}$ implies $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \mu = 0, \forall t \geq 2$. Then
\[
\sum_{t=1}^T \left( \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \mu \right) \|\bar{x}^t - x^*\|^2 = \left( \frac{1}{\eta_1} - \mu \right) \|\bar{x}^1 - x^*\|^2 \leq \mu cD^2. 
\]
By substituting (38) into (36), we derive (30).

\[

\]

The following key lemma analyzes the relationship between the projection error, the consensus error, and the compressor error, and makes it possible to control these errors by the consensus stepsize $\bar{\gamma}$ and the gradient descent stepsize $\eta_t$. 

\[

\]
Lemma 2. Suppose Assumptions 1 and 4 hold. Consider Algorithm 1 with the consensus stepsize $\gamma \in (0, 1)$ and arbitrary gradient descent stepsizes $\{\eta_t\}_{t=1}^T$.

(i) $\mathbb{E}_Q \|X^{t+1} - \tilde{X}^{t+1}\|^2 \leq 2(1 - \omega)\beta^2\gamma^2 \|X^t - \tilde{X}^t\|^2 + 2NG^2\eta_t^2$. \hspace{1cm} (39)

(ii) $\mathbb{E}_Q \|X^{t+1} - \tilde{X}^{t+1}\|^2 \leq (1 - \gamma)\mathbb{E}_Q \|X^t - \tilde{X}^t\|^2 + 9 \left(1 + \frac{2}{\gamma}\right)NG^2\eta_t^2 + 9 \left(1 + \frac{2}{\omega}\right)(1 - \omega)^2\gamma^2 \|X^t - \tilde{X}^t\|^2$. \hspace{1cm} (40)

(iii) $\mathbb{E}_Q \|X^{t+1} - \tilde{X}^{t+1}\|^2 \leq 3 \left(1 + \frac{2}{\gamma}\right)\beta^2\mathbb{E}_Q \|X^t - \tilde{X}^t\|^2 + 9 \left(1 + \frac{2}{\omega}\right)(1 - \omega)^2\beta^2 \|X^t - \tilde{X}^t\|^2$. \hspace{1cm} (41)

Proof. First of all, by Assumption 2 and the update rule of $X^{t+1}$, we have

$$\mathbb{E}_Q \|X^t - \tilde{X}^t\|^2 \leq (1 - \omega)\|X^t - \tilde{X}^t\|^2. \hspace{1cm} (42)$$

(i) By Assumption 1, $\sum_{j \in \mathcal{N}} a_{ij} x_j^t \in \mathcal{K}$ and $1 - \gamma x_i^t + \gamma \sum_{j \in \mathcal{N}_i} a_{ij} x_j^t - \bar{a}_i^{t+1}$ by the projection property (ii),

$$\|x_i^{t+1}\| = \|P_{\mathcal{K}}(\bar{x}_i^{t+1}) - \bar{a}_i^{t+1}\| \leq (1 - \gamma) x_i^t + \gamma \sum_{j \in \mathcal{N}_i} a_{ij} x_j^t - \bar{a}_i^{t+1} = \left| x_i^t + \gamma \sum_{j \in \mathcal{N}_i} a_{ij} x_j^t - \gamma \sum_{j \in \mathcal{N}_i} a_{ij} \bar{x}_j^{t+1} - \eta_t \nabla f_i^t(x_i^t) \right| = -\gamma L_i \left( X^t - \tilde{X}^t \right) + \eta_t \nabla f_i^t(x_i^t).$$

Then, we can estimate the total projection error as

$$\|R^{t+1}\|^2 = \sum_{i=1}^N \|x_i^{t+1}\|^2 \leq \sum_{i=1}^N \left( 2 \gamma L_i \left( X^t - \tilde{X}^{t+1} \right) \right)^2 + 2\|\eta_t \nabla f_i^t(x_i^t)\|^2 \leq 2 \gamma L_i \left( X^t - \tilde{X}^{t+1} \right)^2 + 2NG^2\eta_t^2 \leq 2NG^2\eta_t^2 + 2NG^2\eta_t^2, \hspace{1cm} (43)$$

which is controlled by $\gamma$ and $\eta_t$. By taking expectation over the internal randomness of the compressor $Q$ with respect to the above inequality and using (42), we derive (39).

(ii) Under Assumption 1, $ML = LM = 0$ and $L \tilde{X}^t = LMX^t = 0$. By the update rule of $X^{t+1}$,

$$\|X^{t+1} - \tilde{X}^{t+1}\|^2 \leq \left(1 + \frac{\gamma\delta}{2}\right) \|\tilde{X}^t - \tilde{X}^t\|^2 + \left(1 + \frac{2}{\gamma\delta}\right) \|\gamma L \tilde{X}^{t+1} - \eta_t \nabla F^t(X^t) + R^{t+1}\|^2.$$

$$= \left(1 + \frac{\gamma\delta}{2}\right) \|\tilde{X}^t - \tilde{X}^t\|^2 + \left(1 + \frac{2}{\gamma\delta}\right) \left(1 - \gamma L \right) \left( X^t - \tilde{X}^t \right) - \gamma L \left( \tilde{X}^{t+1} - X^t \right) - \eta_t \left( I - M \right) \nabla F^t(X^t) + (I - M) R^{t+1}\|^2 \hspace{1cm} (44)$$
The first term can be estimated by
\[ \| (I - \gamma L) (X^t - \tilde{X}^t) \| = \| (1 - \gamma) I + \gamma A \| (X^t - \tilde{X}^t) \| \]
\[ = (1 - \gamma) \| X^t - \tilde{X}^t \| + \gamma \| (A - M) (X^t - \tilde{X}^t) \| \]
\[ \leq (1 - \gamma) \| X^t - \tilde{X}^t \| + \gamma (1 - \delta) \| X^t - \tilde{X}^t \| \]
\[ = (1 - \gamma \delta) \| X^t - \tilde{X}^t \| , \] (45)
because \( M (X^t - \tilde{X}^t) = \tilde{X}^t - \tilde{X}^t = 0 \) and \( \| A - M \|_2 = 1 - \delta \). The expectation of the second term can be estimated by
\[ \mathbb{E}_Q \left( -\gamma L \left( \tilde{X}^{t+1} - X^t \right) - \eta_t (I - M) \nabla F^t(X^t) + (I - M) R^{t+1} \right)^2 \]
\[ \leq \mathbb{E}_Q \left( 3 \| \gamma L (\tilde{X}^{t+1} - X^t) \|^2 + 3 \| \eta_t (I - M) \nabla F^t(X^t) \|^2 + 3 \| (I - M) R^{t+1} \|^2 \right) \]
\[ \leq 3 \left( \gamma^2 \beta^2 (1 - \omega) \| \tilde{X}^t - X^t \|^2 + NG^2 \eta_t^2 + 2(1 - \omega) \beta^2 \gamma^2 \| X^t - \tilde{X}^t \|^2 + 2NG^2 \eta_t^2 \right) \]
\[ = 9 \left( 1 - \omega \right) \beta^2 \gamma^2 \| X^t - \tilde{X}^t \|^2 + NG^2 \eta_t^2 \] (46)

By taking expectation over \( Q \) w.r.t the inequality (44), together with (45) and (46), we obtain
\[ \mathbb{E}_Q \| X^{t+1} - \tilde{X}^{t+1} \|^2 \leq \left( 1 + \frac{\gamma \delta}{2} \right) (1 - \gamma \delta)^2 \mathbb{E}_Q \| X^t - \tilde{X}^t \|^2 + 9 \left( 1 + \frac{2}{\gamma \delta} \right) NG^2 \eta_t^2 \]
\[ + 9 \left( 1 + \frac{2}{\gamma \delta} \right) (1 - \omega) \beta^2 \gamma \| X^t - \tilde{X}^t \|^2 \] (47)
\[ \leq (1 - \gamma \delta) \mathbb{E}_Q \| X^t - \tilde{X}^t \|^2 + 9 \left( 1 + \frac{2}{\gamma \delta} \right) NG^2 \eta_t^2 \]
\[ + 9 \left( 1 + \frac{2}{\gamma \delta} \right) (1 - \omega) \beta^2 \gamma \| X^t - \tilde{X}^t \|^2 , \] (48)

since \( \left( 1 + 2 \delta \right) (1 - \gamma \delta)^2 \leq \left( 1 - 2 \delta \right) (1 - \gamma \delta) \leq 1 - \gamma \delta \) and \( \gamma \leq 1 \).

(iii) Similarly to the procedure of (ii), we have
\[ \| X^{t+1} - \tilde{X}^{t+1} \|^2 \leq \| (I + \gamma L) (X^t - \tilde{X}^t) - \eta_t \nabla F^t(X^t) + R^{t+1} - \tilde{X}^{t+1} \|^2 \]
\[ = \| (I + \alpha) L \| (X^t - \tilde{X}^t) - \eta_t \nabla F^t(X^t) + R^{t+1} \|^2 \]
\[ \leq \left( 1 + \frac{\omega}{2} \right) \| (I + \gamma L) (X^t - \tilde{X}^t) \|^2 \]
\[ + \left( 1 + \frac{2}{\omega} \right) \| -\gamma L (X^t - \tilde{X}^t) - \eta_t \nabla F^t(X^t) + R^{t+1} \|^2 . \] (49)
The expectation of the first term can be estimated by
\[ \mathbb{E}_Q \| (I + \gamma L) (X^t - \tilde{X}^t) \|^2 \leq (1 + \gamma \beta)^2 \mathbb{E}_Q \| X^t - \tilde{X}^t \|^2 \]
\[ \leq (1 + \gamma \beta)^2 (1 - \omega) \| X^t - \tilde{X}^t \|^2 , \] (50)
due to \( \| I + \gamma L \|_2 = 1 + \gamma \| L \|_2 = 1 + \gamma \beta \), since the eigenvalues of \( \gamma L \) are positive. The expectation of the second term can be estimated by
\[ \mathbb{E}_Q \| -\gamma L (X^t - \tilde{X}^t) - \eta_t \nabla F^t(X^t) + R^{t+1} \|^2 \]
\[ \leq \mathbb{E}_Q \left( 3 \| \gamma L (X^t - \tilde{X}^t) \|^2 + 3 \| \eta_t \nabla F^t(X^t) \|^2 + 3 \| R^{t+1} \|^2 \right) \]
\[ \leq 3 \gamma^2 \beta^2 \mathbb{E}_Q \| X^t - \tilde{X}^t \|^2 + 6NG^2 \eta_t^2 \] (51)
By taking expectation over $Q$ w.r.t the inequality (49), together with (50) and (51), we obtain
\[
E_Q \left\| X^{t+1} - \hat{X}^{t+1} \right\|^2 \leq 3 \left( 1 + \frac{2}{\omega} \right) \gamma^2 \beta^2 E_Q \left\| X^t - \hat{X}^t \right\|^2 + 9 \left( 1 + \frac{2}{\omega} \right) NG^2 \eta^2_t
\]
\[+ \left( \left( 1 + \frac{\omega}{\omega} \right) (1 - \omega) (1 + \gamma \beta)^2 + 6 \left( 1 + \frac{2}{\omega} \right) (1 - \omega) \beta^2 \gamma^2 \right) \left\| X^t - \hat{X}^t \right\|^2 \]
\[\leq 3 \left( 1 + \frac{2}{\omega} \right) \gamma^2 \beta^2 E_Q \left\| X^t - \hat{X}^t \right\|^2 + 9 \left( 1 + \frac{2}{\omega} \right) NG^2 \eta^2_t
\]
\[+ \left( \left( 1 + \frac{\omega}{\omega} \right) (1 - \omega) (1 + (\beta^2 + 2\beta) \gamma)^2 + 6 \left( 1 + \frac{2}{\omega} \right) (1 - \omega) \beta^2 \gamma^2 \right) \left\| X^t - \hat{X}^t \right\|^2 , \quad (52)
\]
since $\gamma^2 \leq \gamma$ for $\gamma \in (0, 1]$.

Lemma 3. Suppose Assumptions[2][2] and[4] hold. Consider Algorithm[7] with the consensus stepsize $\gamma$ chosen as (5) and arbitrary gradient descent stepsizes $\{\eta_t\}_{t=1}$. Define
\[
e_t := \left\| \frac{E_Q \left\| X^{t+1} - \hat{X}^{t+1} \right\|}{E_Q \left\| X^t - \hat{X}^t \right\|} \right\|.
\]
Then for $t = 1, \cdots, T$,
\[
e_t + 1 \leq \left( 1 - \frac{3}{4} \delta \gamma \right) e_t + 18 \left( 1 + \frac{1}{\gamma^3} + \frac{1}{\omega} \right) NG^2 \eta^2_t . \quad (53)
\]

Proof. By Lemma[2] we have
\[
\left[ \frac{E_Q \left\| X^{t+1} - \hat{X}^{t+1} \right\|}{E_Q \left\| X^t - \hat{X}^t \right\|} \right] \leq U(\gamma) \left[ \frac{E_Q \left\| X^t - \hat{X}^t \right\|}{E_Q \left\| X^t - \hat{X}^t \right\|} \right] + 9NG^2 \eta^2_t \left[ 1 + \frac{2}{\gamma^3} \right] , \quad (54)
\]
where
\[
U(\gamma) := \begin{bmatrix}
1 - \delta \gamma & \frac{9}{3} (1 + \frac{1}{2}) (1 - \omega) \beta^2 \gamma \\
\frac{9}{3} (1 + \frac{1}{2}) \beta^2 \gamma & \left( 1 + \frac{\omega}{\omega} \right) (1 - \omega) (1 + (\beta^2 + 2\beta) \gamma) + 6 \left( 1 + \frac{2}{\omega} \right) (1 - \omega) \beta^2 \gamma
\end{bmatrix}.
\]
For notation simplicity, we denote $u_1 = 9 (1 + \frac{1}{2}) (1 - \omega) \beta^2$, $u_2 = 3 \left( 1 + \frac{2}{\omega} \right) \beta^2$, $u_3 = \left( 1 + \frac{\omega}{\omega} \right) (1 - \omega) (\beta^2 + 2\beta) + 6 \left( 1 + \frac{\omega}{\omega} \right) (1 - \omega) \beta^2$, and write
\[
U(\gamma) = \begin{bmatrix}
1 - \delta \gamma & \frac{u_1 \gamma}{u_2 \gamma} \\
\frac{u_1 \gamma}{u_2 \gamma} & 1 - \frac{u_1 \gamma}{u_2 \gamma} + u_3 \gamma
\end{bmatrix}.
\]
By the definition of $e_t$, we obtain
\[
e_{t+1} \leq \left\| U(\gamma) \right\|_2 e_t + 9NG^2 \eta^2_t \left[ 1 + \frac{2}{\gamma^3} \right] \left[ 1 + \frac{\omega}{\omega} \right] \]
\[\leq \rho(U(\gamma)) e_t + 9NG^2 \eta^2_t \left[ 1 + \frac{2}{\gamma^3} + 1 + \frac{2}{\omega} \right] . \quad (55)
\]
Next, we focus on the spectrum radius of the matrix $U(\gamma)$. The characteristic polynomial of $U(\gamma)$ is
\[
h(\tau) = \det (\tau I - U(\gamma)) = \tau^2 - \left( 1 - \delta \gamma + 1 - \omega - \frac{\omega^2}{2} + u_3 \gamma \right) \tau + (1 - \delta \gamma) \cdot \left( 1 - \omega - \frac{\omega^2}{2} + u_3 \gamma \right) - u_1 u_2 \gamma^2.
\]
Since
\[
\Delta = \left( 1 - \delta \gamma + 1 - \omega - \frac{\omega^2}{2} + u_3 \gamma \right)^2 - 4 \left( (1 - \delta \gamma) \cdot \left( 1 - \omega - \frac{\omega^2}{2} + u_3 \gamma \right) - u_1 u_2 \gamma^2 \right)
\]
\[= \left( 1 - \delta \gamma - \left( 1 - \frac{\omega}{2} - \frac{\omega^2}{2} + u_3 \gamma \right) \right)^2 + 4u_1 u_2 \gamma^2 \geq 0 , \quad (56)
\]
the equation \( h(\tau) = 0 \) has two roots \( \tau_1 \) and \( \tau_2 \). Since \( 1 - \delta \gamma + 1 - \frac{\omega^2}{2} + u_3 \gamma \geq 0 \), \(\rho(U(\gamma)) = \max\{\tau_1, \tau_2\} = \frac{1}{2} \left( 1 - \delta \gamma + 1 - \frac{\omega^2}{2} + u_3 \gamma \right)^2. \) (57)

When \( \gamma \leq \frac{2\delta(\omega^2 + \omega)}{16u_1u_2 + 4u_3\delta + 3\delta^2}, \) (58)

it can be verified that \( \Delta \leq \left( 1 - \frac{\gamma \delta}{2} - \left( 1 - \frac{\omega^2}{2} + u_3 \gamma \right)^2, \right) \) and then, \( \rho(U(\gamma)) \leq \frac{1}{2} \left( 1 - \frac{\gamma \delta}{2} + \left( 1 - \frac{\omega^2}{2} + u_3 \gamma \right) + \left( 1 - \frac{\gamma \delta}{2} - \left( 1 - \frac{\omega^2}{2} + u_3 \gamma \right) \right) \right) \)

\( = 1 - \frac{3}{4} \gamma \delta. \) (60)

We take \( \gamma = \gamma(\omega) := \frac{3\delta}{4} \frac{2\delta(\omega^2 + \omega)}{16u_1u_2 + 4u_3\delta + 3\delta^2} \)

\( = \frac{3\delta^2(\omega + 1)}{48(\delta^2 + 18\delta^2 + 36\delta^2)(\omega + 2)(1 - \omega) + 4\delta^2(\delta^2 + \delta)(\omega + 2)(1 - \omega)\omega + 6\delta^2\omega}, \) (61)

which satisfies (58) since \( \frac{3\delta}{4} \leq 1. \) Notice that \( \gamma(\omega) \) increases monotonically with \( \omega, \) and \( \gamma(0) = 0, \gamma(1) = 1. \) Thus, \( \gamma(\omega) \in (0, 1] \) for \( \omega \in (0, 1], \) which meets the algorithm design requirement. Then, the lemma is proved.

Lemma 4. Let \( \{e_t\}_{t \geq 1} \) denotes a sequence of real values satisfying \( e_1 = 0 \) and

\[ e_{t+1} \leq (1 - p)e_t + q \eta_t^2, \] (62)

for parameters \( p \in (0, 1), q > 0, \) and the stepsize sequence \( \{\eta_t\}_{t \geq 1} \) satisfying either of the following conditions

(i) \( \eta_t = \frac{b}{\sqrt{t + c}} \) for constants \( c \geq \frac{2}{p}, b \geq 0, \)

(ii) \( \eta_t = \frac{b}{e^{t+c}} \) for constants \( c \geq \frac{4}{p}, b \geq 0. \)

Then for any \( t \geq 1, \)

\[ e_t \leq \frac{2q}{p} \eta_t^2. \] (63)

Proof. We proceed the proof by induction. For \( t = 1, \) the statement holds since \( e_1 = 0. \) Suppose that the statement holds for \( t. \) Then for \( t + 1, \)

\[ e_{t+1} \leq (1 - p)e_t + q \eta_t^2 \leq (1 - p)\frac{2q}{p} \eta_t^2 + q \eta_t^2. \] (64)

It remains to prove

\[ (1 - p)\frac{2q}{p} \eta_t^2 + q \eta_t^2 \leq \frac{2q}{p} \eta_{t+1}^2. \] (65)

As for the condition (i),

\[ \frac{\eta_{t+1}^2}{\eta_t^2} = \frac{t + c}{t + c + 1} = 1 - \frac{1}{t + c + 1} > 1 - \frac{p}{2}, \quad \forall t \geq 1. \] (66)

As for the condition (ii),

\[ \frac{\eta_{t+1}^2}{\eta_t^2} = \left( \frac{t + c}{t + c + 1} \right)^2 > 1 - \frac{2}{t + c + 1} > 1 - \frac{p}{2}, \quad \forall t \geq 1. \] (67)

Thus, the conclusion follows.
Proof of Theorem 1

By Lemma 3 and Lemma 4, we have

\[ e_t \leq \frac{48}{\gamma \delta} \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) N G^2 \eta_t^2. \]  

(68)

According to the Jensen’s Inequality,

\[ \mathbb{E}_Q \|X^t - \hat{X}^t\|^2 \leq \sqrt{\mathbb{E}_Q \|X^t - \hat{X}^t\|^2} \leq \sqrt{\frac{48}{\gamma \delta} \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) N G^2 \eta_t^2} \]

\[ \leq 4\sqrt{3} \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) \sqrt{N G\eta_t}. \]  

(69)

Similarly to the procedure of (44), we estimate

\[ \|\hat{X}^t - \hat{X}^t+1\|^2 = \|X^t - \gamma L \hat{X}^t+1 - \eta_t \nabla F^t(X^t) - \hat{X}^t\|^2 \]

\[ = \|(I - \gamma L) (X^t - \hat{X}^t) - \gamma L (\hat{X}^t+1 - X^t) - \eta_t \nabla F^t(X^t)\|^2 \]

\[ \leq \left( 1 + \frac{\gamma \delta}{2} \right) \|(I - \gamma L) (X^t - \hat{X}^t)\|^2 \]

\[ + \left( 1 + \frac{2}{\gamma \delta} \right) \left( 1 + 2 \right) \|\gamma L (\hat{X}^t+1 - X^t)\|^2 + \left( 1 + \frac{1}{2} \right) \|\eta_t \nabla F^t(X^t)\|^2 \].  

(70)

Together with the estimate of \( \mathbb{E}_Q \|R_{t+1}\|^2 \) and the choice of \( \gamma \) in (5), we have

\[ \mathbb{E}_Q \left( \|\hat{X}^t - \hat{X}^t+1\|^2 + 3\|R_{t+1}\|^2 \right) \]

\[ \leq \left( 1 + \frac{\gamma \delta}{2} \right) (1 - \gamma \delta)^2 \mathbb{E}_Q \|X^t - \hat{X}^t\|^2 + 3 \left( 1 + \frac{2}{\gamma \delta} \right) (1 - \omega)\beta^2 \gamma^2 \|X^t - \hat{X}^t\|^2 \]

\[ + \frac{3}{2} \left( 1 + \frac{2}{\gamma \delta} \right) N G^2 \eta_t^2 + 3 \left( 2(1 - \omega)\beta^2 \gamma^2 \|X^t - \hat{X}^t\|^2 + 2 N G^2 \eta_t^2 \right) \]

\[ \leq (1 - \gamma \delta) \mathbb{E}_Q \|X^t - \hat{X}^t\|^2 + 9 \left( 1 + \frac{2}{\delta} \right) (1 - \omega)\beta^2 \gamma \|X^t - \hat{X}^t\|^2 + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2 \]

\[ = [1 \ 0] U(\gamma) \left[ \mathbb{E}_Q \|X^t - \hat{X}^t\|^2 \right] + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2 \]

\[ \leq \rho(U(\gamma)) e_t + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2 \]

\[ \leq \left( 1 - \frac{3}{4} \gamma \delta \right) \frac{48}{\gamma \delta} \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) N G^2 \eta_t^2 + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2 \]

\[ = \left( \frac{48}{\gamma \delta} - 36 \right) \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2, \]  

(71)

where \( U(\gamma) \) is defined in Lemma 3

For the strongly convex case (ii), we substitute (69) and (71) into (30), and derive

\[ \mathbb{E}_Q R(j, T) \leq \mu c D^2 + N G^2 \sum_{t=1}^{T} \eta_t + (2 \sqrt{N} + N)G \sum_{t=1}^{T} 4 \sqrt{3} \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) \sqrt{N G} \eta_t \]

\[ + \sum_{t=1}^{T} \frac{1}{2} \left( \frac{48}{\gamma \delta} - 36 \right) \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) + \left( \frac{15}{2} + \frac{3}{\gamma \delta} \right) N G^2 \eta_t^2 \]

\[ \leq \mu c D^2 + 4 \sqrt{3} \left( 2 \sqrt{N} + \frac{2 \sqrt{3}}{\gamma \delta} + 1 \right) \left( 1 + \frac{1}{\gamma \delta} + \frac{1}{\omega} \right) N G^2 \sum_{t=1}^{T} \eta_t, \]  

(72)
where the last inequality holds since \((4\sqrt{3} + \frac{3}{2} - 18) / \gamma \delta < 0\) and \(1 + (4\sqrt{3} - 18)(1 + \frac{1}{\omega}) + \frac{15}{4} < 0\) for \(\omega \in (0, 1]\). Consider the gradient stepsize \(\eta_t = \frac{1}{\mu(t+c)}\) for \(c \geq \frac{16}{\delta \gamma} > 1\), and then,

\[
\sum_{t=1}^{T} \eta_t = \sum_{t=1}^{T} \frac{1}{\mu(t+c)} \leq \frac{1}{\mu} \int_{0}^{T} \frac{1}{s+c} ds = \frac{1}{\mu} \ln(s+c)|_{0}^{T} \leq \frac{1}{\mu} \ln(T+c).
\]

Combining (72) with (73), we obtain (7).

For the convex case (i), the gradient stepsize \(\eta_t = \frac{D}{G \sqrt{T+c}}\), and then,

\[
\sum_{t=1}^{T} \eta_t = \sum_{t=1}^{T} \frac{D}{G \sqrt{T+c}} \leq \frac{D}{G} \int_{0}^{T} \frac{1}{\sqrt{s+c}} ds = \frac{D}{G} 2\sqrt{\sqrt{T+c}} \leq \frac{2D}{G} \sqrt{T+c}.
\]

(74)

By substituting (69), (71), and (74) into (29), we derive

\[
\mathbb{E}_{Q} R(j, T) \leq \frac{N D^2}{2 \eta_T} + 4\sqrt{3} \left(\sqrt{N} + \frac{2\sqrt{3}}{\gamma \delta} + 1\right) \left(1 + \frac{1}{\gamma \delta} + \frac{1}{\omega}\right) \frac{G^2}{2} \sum_{t=1}^{T} \eta_t
\]

\[
\leq \frac{N D^2 G \sqrt{T+c}}{2} + 4\sqrt{3} \left(\sqrt{N} + \frac{2\sqrt{3}}{\gamma \delta} + 1\right) \left(1 + \frac{1}{\gamma \delta} + \frac{1}{\omega}\right) \frac{G^2}{2} \frac{2D}{G} \sqrt{T+c}.
\]

Then the theorem is proved.

\[\square\]

\section{Proofs of Section 3}

Algorithm 2 actually performs the gradient descent scheme on the function \(\hat{f}_i^j(x) = \mathbb{E}_{u \in B}[f_i^j(x + \epsilon u)]\) restricted to the convex set \((1 - \zeta)\mathcal{K}\). By Assumptions 6 and 7, as well as the construction of \(g_i^j\),

\[
\| \nabla \hat{f}_i^j(x) \| = \| \mathbb{E}[g_i^j] \| \leq \mathbb{E}[\|g_i^j\|] \leq \mathbb{E}\left[ \frac{d}{\epsilon} \| f_i^j \| \| u_i^j \| \right] \leq \frac{d B}{\epsilon} := G, \quad \forall i \in \mathcal{V}, t = 1, \cdots, T,
\]

\[
\| x - y \| \leq 2R := D, \quad \forall x, y \in (1 - \zeta)\mathcal{K}.
\]

The remaining gaps include 1) the difference between the case of the loss function \(f_i^j\) and that of \(\hat{f}_i^j\); 2) the difference between the case of the feasible set \((1 - \zeta)\mathcal{K}\) and that of \(\mathcal{K}\). As for 1), by Assumption 7 we have

\[
\| f_i^j(x) - \hat{f}_i^j(x) \| = \| \mathbb{E}_{u}[f_i^j(x + \epsilon u)] - \hat{f}_i^j(x) \| \leq \mathbb{E}_{u} \| f_i^j(x + \epsilon u) - \hat{f}_i^j(x) \| \leq \epsilon l_c,
\]

and thus,

\[
\hat{f}_i^j(x) - \epsilon l_c \leq f_i^j(x) \leq \hat{f}_i^j(x) + \epsilon l_c.
\]

As for 2), we have the following lemma from [4].

\textbf{Lemma 5.} The optimum in \((1 - \zeta)\mathcal{K}\) is near the optimum in \(\mathcal{K}\).

\[
\min_{x \in (1 - \zeta)\mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x) \leq 2\zeta BNT + \min_{x \in \mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x).
\]

(76)

By Lemma 5, we can obtain the regret bounds in the one-point bandit setting upon the obtained results in the full information setting.

\[
\mathbb{E}[R(j, T)] = \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x_i^j) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x)
\]

\[
\leq \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x_i^j) - \min_{x \in (1 - \zeta)\mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^j(x) + 2\zeta BNT
\]

\[
\leq \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} \left( \hat{f}_i^j(x_i^j) + \epsilon l_c \right) - \min_{x \in (1 - \zeta)\mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} \left( \hat{f}_i^j(x_i^j) - \epsilon l_c \right) + 2\zeta BNT
\]

\[
= \mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{f}_i^j(x_i^j) - \min_{x \in (1 - \zeta)\mathcal{K}} \sum_{t=1}^{T} \sum_{i=1}^{N} \hat{f}_i^j(x_i^j) + 2\epsilon NT + 2\zeta BNT.
\]

(77)
Proof of Theorem\textsuperscript{2}

(i) (Convex case) From Theorem\textsuperscript{1} part (i), we have
\[
\mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x_i^t) - \min_{x \in (1-\zeta)K} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x) \leq \left( \frac{1}{2} + 2H \right) N \frac{dB}{\epsilon} 2R\sqrt{T} + c,
\]
where $H$ is defined in [3]. Then by (77) with $\zeta = \frac{\epsilon}{T}$,
\[
\mathbb{E} [R(j, T)] \leq (1 + 4H) N \frac{dB R}{\epsilon} \sqrt{T} + c + 2\epsilon NT + 2 \frac{\epsilon^2}{r} BNT.
\]
We choose $\epsilon = \left( \frac{(1 + 4H) dB R}{2(1 + \frac{\epsilon}{T})} \right)^{\frac{T}{4}}$ to minimize the right hand of the above inequality and then obtain the conclusion.

(ii) (Strongly convex case) From Theorem\textsuperscript{1} part (ii), we have
\[
\mathbb{E} \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t(x_i^t) - \min_{x \in (1-\zeta)K} \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t(x) \leq 4\mu c R^2 + H \frac{N \mu^2 d^2 B^2}{\mu c^2} \ln(T + c),
\]
where $H$ is defined in [3]. Then by (77) with $\zeta = \frac{\epsilon}{T}$,
\[
\mathbb{E} [R(j, T)] \leq 4\mu c R^2 + H \frac{N \mu^2 d^2 B^2}{\mu c^2} \ln(T + c) + 2\epsilon NT + 2 \frac{\epsilon^2}{r} BNT.
\]
We choose $\epsilon = \left( \frac{N \mu^2 d^2 B^2}{\mu c^2} \ln(T + c)}{2(1 + \frac{\epsilon}{T})} \right)^{\frac{T}{4}}$ to minimize the right hand of the above inequality and then obtain the conclusion.

\[\square\]

D Proofs of Section\textsuperscript{4}

The proof in the two-point bandit case takes a similar procedure as that in the one-point case. By Assumptions\textsuperscript{6} and\textsuperscript{7} as well as the construction of $g_i^t$,
\[
\|\nabla \tilde{f}_i^t(x)\| = \|E [g_i^t]\| \leq \|g_i^t\| \leq E \left[ \frac{d}{2\epsilon} \|f_i^t(x_i^t + \epsilon u_i^t) - f_i^t(x_i^t - \epsilon u_i^t)\| \|u_i^t\| \right] \leq \frac{d}{2\epsilon} \|u_i^t\|^2 = dl := G, \quad \forall i \in V, t = 1, \cdots, T,
\]
\[
\|x - y\| \leq 2R := D, \quad \forall x, y \in (1-\zeta)K.
\]

Similar to the Lemma 2 in [5], we have

Lemma 6. For any point $x \in K$,
\[
\sum_{t=1}^{T} \sum_{i=1}^{N} \frac{f_i^t(y_{i,1}^t) + f_i^t(y_{i,2}^t)}{2} - \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x) \leq \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t(x_i^t) - \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t((1-\zeta)x) + 3NTG \epsilon + NTGD\zeta.
\]

By Lemma\textsuperscript{6} for $x^* = \arg\min_{x \in K} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x)$,
\[
\mathbb{E} [R_2(j, T)] = E \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{f_i^t(y_{i,1}^t) + f_i^t(y_{i,2}^t)}{2} - \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x^*) \leq \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t(x_i^t) - \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t((1-\zeta)x^*) + 3NTG \epsilon + NTGD\zeta
\]
\[
\leq \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{f}_i^t(x_i^t) - \min_{x \in (1-\zeta)K} \sum_{t=1}^{T} \sum_{i=1}^{N} f_i^t(x) + 3NTd \epsilon + 2NTdR\zeta
\]
We choose $\epsilon$.

Also, for strongly convex losses, the gradient descent stepsize can be written as $\eta = H \frac{T}{c}$ where $H$ is defined in (8). Then by (83) with $\zeta = \frac{c}{r}$,

$$E[R_2(j, T)] \leq (1 + 4H) NdR\sqrt{T} + \epsilon + \left(3 + \frac{2R}{r}\right) NdT\epsilon.$$  

We choose $\epsilon = \frac{1}{\sqrt{T}}$ and then obtain the conclusion.

(ii) (Strongly convex case) From Theorem 1 part (ii), we have

$$E\left[ R_2(j, T) \right] \leq (1 + 4H) NdR\sqrt{T} + \epsilon + \left(3 + \frac{2R}{r}\right) NdT\epsilon.$$  

We choose $\epsilon = \frac{\ln(T)}{T}$ and then obtain the conclusion.

### E Parameters selection details

The theoretical value of the consensus stepsize $\gamma$ depends on the compression ratio $\omega$ and the graph parameters $\delta, \beta$, which is pretty conservative. We tune $\gamma$ for each experiment. The gradient descent stepsizes of DC-DOGD, DC-DOBD and DC-DO2BD for convex losses can be written in a unified form as $\eta_t = \frac{b}{\sqrt{T}}$, where $c = \frac{N}{3\gamma \delta}$ as in theorems and $b$ is tuned from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. Also, for strongly convex losses, the gradient descent stepsize can be written as $\eta_t = \frac{b}{\sqrt{T}}$, where $c = \frac{10}{\sqrt{T}}\gamma$ as in theorems and the exploration parameter $\epsilon$ is tuned from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. For each experiment, $b$ and $\epsilon$ are tuned by grid search.

#### Parameters in Fig. 1

In this experiment, we set $\gamma = 0.26$ for QSGD with $\omega = 0.3$ over $G(9, 18)$. The parameters $b$ and $\epsilon$ for the proposed algorithms are given in Table 2. The parameters for ECD-AMDGrad are chosen as suggested in [6].

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Convex losses</th>
<th>Strongly convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC-DOGD</td>
<td>$b$ = 0.1 \ $\epsilon$ = \</td>
<td>$b$ = 1 \ $\epsilon$ = \</td>
</tr>
<tr>
<td>DC-DOBD</td>
<td>$b$ = 0.01 \ $\epsilon$ = 0.5</td>
<td>$b$ = 0.05 \ $\epsilon$ = 0.5</td>
</tr>
<tr>
<td>DC-DO2BD</td>
<td>$b$ = 0.1 \ $\epsilon$ = 0.05</td>
<td>$b$ = 0.5 \ $\epsilon$ = 0.01</td>
</tr>
</tbody>
</table>

#### Parameters in Fig. 2

When studying the impacts of compression ratio and compressor type, we take DC-DOGD with strongly convex losses over the graph $G(N, 2N)$ as an example, and set $b = 1$. The corresponding $\gamma$ for different compression ratios $\omega$ (with the same compressor type $Top_k$) are given in Table 3 and the corresponding $\gamma$ for different compressor types (with the same compression ratio $\omega = 0.3$) are given in Table 4. DAOL takes the same gradient descent stepsizes as DC-DOGD.

#### Parameters in Fig. 3

When studying the impact of network topology, we take DC-DOGD with strongly convex losses as an example, and set $b = 1$. For the compressor $Top_1$ with $\omega = 0.05$, we set $\gamma = 0.09$. When studying the impact of node number, we take the compressor $Top_2$ with $\omega = 0.1$ and set $\gamma = 0.1$. The parameters $b$ and $\epsilon$ are chosen the same as Table 3.
Table 3: Corresponding $\gamma$ for different $\omega$

<table>
<thead>
<tr>
<th>$\omega$</th>
<th>0.05</th>
<th>0.1</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>0.09</td>
<td>0.1</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 4: Corresponding $\gamma$ for different compressors

<table>
<thead>
<tr>
<th>Compressor</th>
<th>RGossip$_p$</th>
<th>Rand$_k$</th>
<th>Top$_k$</th>
<th>GSGD$_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>0.09</td>
<td>0.09</td>
<td>0.28</td>
<td>0.26</td>
</tr>
</tbody>
</table>

F Additional experiments

We give some additional experiments in the convex cases here. Still, we use the dataset `diabetes-binary-BRFSS2015`\(^5\). The communication graph is generated by the tool `NetworkX`\(^6\) and the best solution is obtained by the tool `Logistic Regression`\(^7\). Our code is available at [https://github.com/happy-math/CC-DOCO](https://github.com/happy-math/CC-DOCO).

![Figure 4: The impact of compression ratio $\omega$. Setting: DC-DOGD with the compressor Top$_k$ over $G(9, 18)$ in the convex case. $b = 1$. The corresponding $\gamma$ for different $\omega$ are chosen as in Table 3.](image)

References


\(^5\)Apache 2.0 open source license
\(^6\)Open source 3-clause BSD license
\(^7\)Open source, commercially usable - BSD license
Figure 5: The impact of compression ratio $\omega$. Setting: DC-DOBD with the compressor $\text{Top}_k$ over $G(9, 18)$ in the convex case. $b = 0.01, \epsilon = 1$. The corresponding $\gamma$ for different $w$ are chosen as in Table 3.

Figure 6: The impact of compression ratio $\omega$. Setting: DC-DO2BD with the compressor $\text{Top}_k$ over $G(9, 18)$ in the convex case. $b = 0.1, \epsilon = 0.05$. The corresponding $\gamma$ for different $w$ are chosen as in Table 3.

Figure 7: The impact of compressor type. Setting: DC-DOGD with the compression ratio $\omega = 0.3$ over $G(9, 18)$ in the convex case. $b = 0.1$. The corresponding $\gamma$ for different compressor types are chosen as in Table 4.

Figure 8: The impact of topology. Setting: DC-DOGD with $\text{Top}_1$, $\omega = 0.05$ in the convex case. $b = 0.1, \gamma = 0.09$. 

26