A COMPARISONS WITH JIN ET AL. (2020).

		Jin et al. (2020)	Our work
Gradient	Batch-size	Full-batch (True Gradient) [Eq. (33)]	Mini-batch Stochastic Gradient
	Distribution	Bounded [Theorem 3]	Sub-Gaussian (bounded as a subset) and heavy-tailed
Residual	Weak Signal Strength	$\Delta(M)$ as a function of client number M without explicit form [Theorem 2 and Remark 3].	$\frac{c_0}{\sqrt{M}}$
	Byzantine	Implicit forms without quantitative results [Theorem 7 and Remark 6].	$\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM}$
	Gradient noise	No, since only bounded gradients are considered.	$\mathcal{O}\left(\exp\left(-\frac{n}{2}\right)\right)$ for Sub-Gaussian noise; $\mathcal{O}\left(1/n^{\frac{p'}{2}}\right)$ $(p' \geq 4)$ for heavy-tailed noise.
Differential Privacy		No for sto-sign compressor; flawed arguments for DP-sign compressor.	$d\log(1+\frac{2B}{\beta})$ for arbitrary gradients; $\left(\frac{\Delta_1}{\beta}\right)$ for gradient pairs with bounded l_1 sensitivity Δ_1.
Partial Client Participation		No.	Theoretically and empirically verified, and build adaptive Byzantine adversaries on it.

Table 3: Point-by-point comparisons with Jin et al. (2020).

B ALGORITHMS

Definition 6. *Let $\mathcal{S} \subseteq [M]$ and let $\widehat{u}_m \in \{\pm 1\}^d$ for $m \in \mathcal{S}$.*
(A.1) The mean aggregation rule is defined as $\mathsf{agg}_{avg}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right) = \frac{1}{|\mathcal{S}|}\sum_{m\in\mathcal{S}} \widehat{u}_m$.
(A.2) The coordinate-wise k-trimmed-mean aggregation rule, denoted by $\mathsf{agg}_{trimmed,k}$, takes $k \in \mathbb{N}$ and $\{\widehat{u}_m,\ m \in \mathcal{S}\}$ as inputs and aggregates each coordinate $i \in [d]$ as follows: (1) sort $\{\widehat{u}_{mi},\ m \in \mathcal{S}\}$, where $\widehat{u}_{mi}$ is the i-th coordinate of $\widehat{u}_m$, in an increasing order; (2) remove the top and bottom k values, and denote the remained clients w. r. t. coordinate i as $\mathcal{R}_i$; (3) if $\mathcal{R}_i = \emptyset$, then $\mathsf{agg}_{trimmed,k,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right) = 1$ (or -1) uniformly at random; otherwise, $\mathsf{agg}_{trimmed,k,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right) = \frac{1}{|\mathcal{R}_i|}\sum_{m\in\mathcal{R}_i} \widehat{u}_m$.
(A.3) The coordinate-wise median aggregation rule, denoted by agg_{median}, aggregates each coordinate $i \in [d]$ as follows: it first sorts the $\{\widehat{u}_{mi},\ m \in \mathcal{S}\}$ in an increasing order. If $|\mathcal{S}|$ is even, $\mathsf{agg}_{median,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right)$ outputs the average of the elements whose ranks are $\frac{|\mathcal{S}|}{2}$ and $\frac{|\mathcal{S}|}{2}+1$. Otherwise, $\mathsf{agg}_{median,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right)$ outputs the element whose rank is $\lceil\frac{|\mathcal{S}|}{2}\rceil$.
(A.4) The coordinate-wise majority vote aggregation rule, denoted by agg_{maj}, aggregates each coordinate $i \in [d]$ as follows: If there are more 1 than -1 in $\{\widehat{u}_{mi},\ m \in \mathcal{S}\}$, then $\mathsf{agg}_{maj,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right)$ outputs 1. If there are more -1 than 1 in $\{\widehat{u}_{mi},\ m \in \mathcal{S}\}$, then $\mathsf{agg}_{maj,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right)$ outputs -1. Otherwise, $\mathsf{agg}_{median,i}\left(\{\widehat{u}_m,\ m \in \mathcal{S}\}\right)$ outputs 0.

C ALTERNATIVE RESULTS

C.1 ALTERNATIVE ASSUMPTIONS

The following two alternative assumptions on the randomness of stochastic gradients are of decreasing levels of stringency.

Assumption 6 (Boundedness). *The ℓ_∞ norm of all possible stochastic gradients is upper bounded. Formally, let $m \in [M]$ be an arbitrary client and $\boldsymbol{g}$ be an arbitrary stochastic gradient that client m obtains. For any coordinate $i \in [d]$, there exists $\widetilde{B}_i > 0$ such that $|g_i| \leq \widetilde{B}_i$. Let $\widetilde{B} = \max_{i\in[d]} \widetilde{B}_i$.*

The following alternative assumption relaxes the boundedness requirement, and allows the stochastic gradients to be supported over the entire $\mathbb{R}^d$.

Assumption 7 (Gaussianity)**.** *For a given client* $m \in [M]$*, at any query* $w \in \mathbb{R}^d$*, the stochastic gradient* $\boldsymbol{g}_m(w)$ *is an independent unbiased estimate of* $\nabla f_m(w)$ *that is coordinate-wise related to the gradient* $\nabla f_m(w)$ *as* $\boldsymbol{g}_{mi}(w) = \nabla f_{mi}(w) + \boldsymbol{\xi}_{mi}\ \forall\, i \in [d]$*,*

where $\boldsymbol{\xi}_{mi} \sim \mathcal{N}\left(0, \sigma_{mi}^2\right)$*. Let* $\sigma^2 := \max_{m\in[M], i\in[d]} \sigma_{mi}^2$*.*

C.2 ALTERNATIVE CONVERGENCE RATES

Corollary 2. *Suppose that Assumptions 3 and 7 hold. Choose* $B = (1+\epsilon_0)B_0$ *for* $\epsilon_0 > \sigma/B_0$ *and* $c_0 = \max\left\{\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}, \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}\right\}$*. Fix* $t \geq 1$ *and* $i \in [d]$*. Let* $c > 0$ *be any given constant such that* $c < \frac{3}{5}$*.*

When the system adversary is adaptive or when the system adversary is static but with $\tau(t) \leq \frac{2}{p^2}\log\frac{6}{c}$*, if* $|\nabla_i F(w(t))| \geq \frac{2(B+\beta)}{pM}\tau(t) + \frac{B+\beta}{2\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) + \frac{c_0}{\sqrt{M}}$*, then Eq.* (4) *holds.*

When the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$*, if* $|\nabla_i F(w(t))| \geq \frac{3(B+\beta)\tau(t)}{M} + \frac{B+\beta}{2\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) + \frac{c_0}{\sqrt{M}}$*, then Eq.* (4) *holds.*

Corollary 3. *Suppose Assumptions 1, 2, 3, and 7 hold. For any given* T*,* $B = (1+\epsilon_0)B_0$ *for* $\epsilon_0 > \frac{\sigma}{B_0}$*, and* c *such that* $0 < c < \frac{3}{5}$*, set the learning rate as* $\eta = \frac{1}{\sqrt{dT}}$ *and* $c_0 := \max\left\{\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}, \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}\right\}$*.*

When the system adversary is adaptive or when the system adversary is static but with $\tau(t) \leq \frac{2}{p^2}\log\frac{6}{c}$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + \frac{d}{\sqrt{2\pi}}(B+\beta)\exp\left(-\frac{n}{2}\right)\right.$$
$$\left. + 2d\frac{c_0}{\sqrt{M}} + 4d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pTM}\right].$$

On the other hand, when the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + \frac{d}{\sqrt{2\pi}}(B+\beta)\exp\left(-\frac{n}{2}\right)\right.$$
$$\left. + 2d\frac{c_0}{\sqrt{M}} + 6d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM}\right].$$

Corollary 4. *Suppose that Assumption 6 holds. Choose* $B = \widetilde{B}$ *and* $c_0 = \max\left\{\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}, \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}\right\}$*. Fix* $t \geq 1$ *and* $i \in [d]$*. Let* $c > 0$ *be any given constant such that* $c < \frac{3}{5}$*.*

When the system adversary is adaptive or when the system adversary is static but with $\tau(t) \leq \frac{2}{p^2}\log\frac{6}{c}$*, if* $|\nabla_i F(w(t))| \geq \frac{2(B+\beta)}{pM}\tau(t) + \frac{c_0}{\sqrt{M}}$*, then Eq.* (4) *holds.*

When the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$*, if* $|\nabla_i F(w(t))| \geq \frac{3(B+\beta)}{M}\tau(t) + \frac{B+\beta}{2\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) + \frac{c_0}{\sqrt{M}}$*, then Eq.* (4) *holds.*

Corollary 5. *Suppose Assumptions 1, 2, and 6 hold. For any given* T *and* c *such that* $0 < c < \frac{3}{5}$*, set the learning rate as* $\eta = \frac{1}{\sqrt{dT}}$ *and* $c_0 := \max\left\{\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}, \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}\right\}$*.*

When the system adversary is adaptive or when the system adversary is static but with $\tau(t) \le \frac{2}{p^2}\log\frac{6}{c}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \le \frac{1}{c}\left[\frac{\left(F(w(0))-F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 4d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pTM}\right].$$

On the other hand, when the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \le \frac{1}{c}\left[\frac{\left(F(w(0))-F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 6d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM}\right].$$

D PROOFS

D.1 AGGREGATION FUNCTIONS

Proof. **[Proof of Proposition 1 (Equivalent to Majority Vote)]** The intuition behind this proof is to show that the signs of all the aggregation rules mentioned in the theorem statement, given $\widehat{u}_m \in \{\pm 1\}^d$ for $m \in \mathcal{S}$, are equivalent to the sign of the k-trimmed-mean aggregation rule.

We first show that for any $k < |\mathcal{S}|/2$, the signs of the outputs of the signs of the aggregation rule $\mathsf{agg}_{\text{trimmed},k}$ are the same. When $k < |\mathcal{S}|/2$, it holds that $\mathcal{R}_i \neq \emptyset$ for each $i \in [d]$. Thus, the aggregation rules $\mathsf{agg}_{\text{trimmed},k}$ with $k < |\mathcal{S}|/2$ is deterministic.

For any given coordinate $i \in [d]$, if the sign of $\mathsf{agg}_{\text{trimmed},k}$ is 0, by definition, we know that there are equal numbers of 1 and -1 in $\{\widehat{u}_{mi} : m \in \mathcal{R}_i\}$, and that the top (resp. bottom) k elements removed from $\{\widehat{u}_{mi} : m \in \mathcal{S}\}$ are 1 (resp. -1). That is, there are equal numbers of 1 and -1 in $\{\widehat{u}_{mi} : m \in \mathcal{S}\}$. Hence, for any $k' \neq k$, as long as the remained set $\mathcal{R}_i'$ after trimming is nonempty (which is ensured by the condition that $k' < |\mathcal{S}|/2$), it holds that $\mathsf{agg}_{\text{trimmed},k'}\left(\{\widehat{u}_{mi} : m \in \mathcal{S}\}\right) = 0$.

If the sign of $\mathsf{agg}_{\text{trimmed},k}$ is -1, we know that there are more -1 than 1 in $\{\widehat{u}_{mi} : m \in \mathcal{R}_i\}$, and that the bottom k elements in $\{\widehat{u}_{mi} : m \in \mathcal{S}\}$ are all -1 whereas the number of 1 in the top k elements is at most k. That is, there are more -1 than 1 in $\{\widehat{u}_{mi} : m \in \mathcal{S}\}$. Hence, we know that for any k', as long as the remained set $\mathcal{R}_i'$ after trimming is nonempty, the sign of $\mathsf{agg}_{\text{trimmed},k'}\left(\{\widehat{u}_{mi} : m \in \mathcal{S}\}\right)$ is -1. Similarly, we can show the case when the sign of $\mathsf{agg}_{\text{trimmed},k}$ is 1.

The above argument, combined with the definition of $\mathsf{agg}_{\text{maj}}$, immediately implies that when $k < |\mathcal{S}|/2$, the signs of $\mathsf{agg}_{\text{trimmed},k}$ and $\mathsf{agg}_{\text{maj}}$ are the same.

Finally, since $\mathsf{agg}_{\text{avg}}$ is $\mathsf{agg}_{\text{trimmed},0}$ and $\mathsf{agg}_{\text{median}} = \mathsf{agg}_{\text{trimmed},\lfloor\frac{|\mathcal{S}|-1}{2}\rfloor}$, the signs of $\mathsf{agg}_{\text{avg}}$, $\mathsf{agg}_{\text{median}}$, and $\mathsf{agg}_{\text{trimmed},k}$ for $k < |\mathcal{S}|/2$ are all the same, proving the theorem.

□

D.2 PRIVACY PRESERVATION

Theorem 5. *(Dwork et al., 2014, Corollary 3.15) Let $\mathcal{M}_i : \mathbb{R}^d \to \{\pm 1\}^d$ be an ϵ_i-differentially private algorithm for $i \in [k]$. Then $\mathcal{M}_{[k]}(x) := (\mathcal{M}_1(x), \cdots, \mathcal{M}_k(x))$ is $\sum_{i=1}^k \epsilon_i$-differentially private.*

Proof of Theorem 1 (Necessity of β). We first consider the setting when $\beta = 0$. Let

$$\mathcal{G} = \{\boldsymbol{g} \in \mathbb{R}^d : \exists i \ s.t. \min\{|\boldsymbol{g}_i - B|, |\boldsymbol{g}_i + B|\} \le 1\}.$$

Let $\boldsymbol{g} \in \mathcal{G}$. Without loss of generality, let us assume that $|\boldsymbol{g}_1 - B| \le 1$, where $\boldsymbol{g}_1$ is the first entry of $\boldsymbol{g}$. If $\boldsymbol{g}_1 \ge B$, then there exists $\boldsymbol{g}' \in \mathbb{R}^d$ such that $\boldsymbol{g}' \neq \boldsymbol{g}$, $\boldsymbol{g}_1' \in (-B, B)$, and $\|\boldsymbol{g} - \boldsymbol{g}'\|_1 \le 1$. Let $\widehat{\boldsymbol{g}}_1$ and $\widehat{\boldsymbol{g}'}_1$ be the compressed values of $\boldsymbol{g}_1$ and $\boldsymbol{g}_1'$ under our compressor in Eq. (2). It holds that

$$\frac{\mathbb{P}\left\{\widehat{\boldsymbol{g}_1'} = -1\right\}}{\mathbb{P}\left\{\widehat{\boldsymbol{g}}_1 = -1\right\}} = \frac{\frac{B - \mathsf{clip}\{\boldsymbol{g}_1', B\}}{2B}}{\frac{B - \mathsf{clip}\{\boldsymbol{g}_1, B\}}{2B}} = \frac{B - \mathsf{clip}\left\{\boldsymbol{g}_1', B\right\}}{B - \mathsf{clip}\left\{\boldsymbol{g}_1, B\right\}} = \frac{B - \mathsf{clip}\left\{\boldsymbol{g}_1', B\right\}}{B - B} = \infty.$$

If $\boldsymbol{g}_1 \in (-B, B)$, then there exists $\boldsymbol{g}' \in \mathbb{R}^d$ such that $\boldsymbol{g}' \neq \boldsymbol{g}$, $\boldsymbol{g}'_1 \geq B$, and $\|\boldsymbol{g} - \boldsymbol{g}'\|_1 \leq 1$. We have

$$\frac{\mathbb{P}\left\{\widehat{\boldsymbol{g}}_1 = -1\right\}}{\mathbb{P}\left\{\widehat{\boldsymbol{g}}'_1 = -1\right\}} = \frac{\frac{B - \mathsf{clip}\{\boldsymbol{g}_1, B\}}{2B}}{\frac{B - \mathsf{clip}\{\boldsymbol{g}'_1, B\}}{2B}} = \frac{B - \mathsf{clip}\left\{\boldsymbol{g}_1, B\right\}}{B - \mathsf{clip}\left\{\boldsymbol{g}'_1, B\right\}} = \frac{B - \mathsf{clip}\left\{g_1, B\right\}}{B - B} = \infty.$$

Since a finite differential privacy quantification does not hold for any pair of gradients $\boldsymbol{g}$ and $\boldsymbol{g}'$, no differential privacy implies as per Definition 1, proving the first part of the theorem.

When $\beta > 0$, for any $\boldsymbol{g}, \boldsymbol{g}' \in \mathbb{R}^d$ such that $\boldsymbol{g}' \neq \boldsymbol{g}$ and $\|\boldsymbol{g} - \boldsymbol{g}'\|_1 \leq 1$, and for each coordinate $i \in [d]$, it holds that

$$\frac{\mathbb{P}\left\{\widehat{\boldsymbol{g}}'_i = -1\right\}}{\mathbb{P}\left\{\widehat{\boldsymbol{g}}_i = -1\right\}} = \frac{\frac{B + \beta - \mathsf{clip}\{\boldsymbol{g}'_1, B\}}{2B + 2\beta}}{\frac{B + \beta - \mathsf{clip}\{\boldsymbol{g}_1, B\}}{2B + 2\beta}} = \frac{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}'_1, B\right\}}{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_1, B\right\}} \leq \frac{2B + \beta}{\beta}.$$

Similarly, we can show the same upper bound for $\mathbb{P}\left\{\widehat{\boldsymbol{g}}'_i = 1\right\} / \mathbb{P}\left\{\widehat{\boldsymbol{g}}_i = 1\right\}$. That is, for the i-th coordinate, the compressor $\mathcal{M}_{B,\beta}$ is coordinate-wise $\log\left(\frac{2B+\beta}{\beta}\right)$- differentially private. By Theorem 5, we conclude that the compressor $\mathcal{M}_{B,\beta}$ is $d \cdot \log\left(\frac{2B+\beta}{\beta}\right)$- differentially private for the entire gradient. □

Proof of Theorem 2 (Smaller Collection of Gradients). For each coordinate $i \in [d]$, it holds that

$$\begin{aligned}
\frac{\mathbb{P}\left\{\widehat{\boldsymbol{g}}'_i = -1\right\}}{\mathbb{P}\left\{\widehat{\boldsymbol{g}}_i = -1\right\}} &= \frac{\frac{B + \beta - \mathsf{clip}\{\boldsymbol{g}'_i, B\}}{2B + 2\beta}}{\frac{B + \beta - \mathsf{clip}\{\boldsymbol{g}_i, B\}}{2B + 2\beta}} = \frac{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}'_i, B\right\}}{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}} \\
&= \frac{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\} + \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\} - \mathsf{clip}\left\{\boldsymbol{g}'_i, B\right\}}{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}} \\
&\leq 1 + \frac{|\boldsymbol{g}_i - \boldsymbol{g}'_i|}{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}} \qquad (7) \\
&\leq 1 + \frac{\Delta_1}{B + \beta - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}} \\
&\leq 1 + \frac{\Delta_1}{\beta + \mathsf{dist}\left(\boldsymbol{g}_i, \mathcal{C}_B\right)}.
\end{aligned}$$

By Theorem 5, we conclude that the compressor $\mathcal{M}_{B,\beta}$ is $\max_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{d} \log\left(1 + \frac{\Delta_1}{\beta + \mathsf{dist}(\boldsymbol{g}_i, \mathcal{C}_B)}\right)$-differentially private for all gradients $\boldsymbol{g} \in \mathcal{G}$. □

Proof of Corollary 1 (Bounded DP with Bounded Sensitivity). By Theorem 2, we conclude that the compressor $\mathcal{M}_{B,\beta}$ is $\max_{\boldsymbol{g} \in \mathcal{G}} \sum_{i=1}^{d} \log\left(1 + \frac{\Delta_1}{\beta + \mathsf{dist}(\boldsymbol{g}_i, \mathcal{C}_B)}\right)$- differentially private for all gradients $\boldsymbol{g} \in \mathcal{G}$. It turns out that this bound can be relaxed, and we start the derivation from Eq. (7):

$$(7) \leq 1 + \frac{|\boldsymbol{g}_i - \boldsymbol{g}'_i|}{\beta}.$$

Now consider the coordinate collection of the gradient pair, by Theorem 5, it remains to bound

$$\begin{aligned}
\sum_{i=1}^{d} \log\left(1 + \frac{|\boldsymbol{g}_i - \boldsymbol{g}'_i|}{\beta}\right) &\leq d \log\left[\frac{1}{d}\sum_{i=1}^{d}\left(1 + \frac{|\boldsymbol{g}_i - \boldsymbol{g}'_i|}{\beta}\right)\right] \quad \text{[Jensen's inequality]} \\
&\leq d \log\left(1 + \frac{\Delta_1}{d\beta}\right) \\
&\leq \frac{\Delta_1}{\beta} \quad \text{[follows from } \log(1 + x) < x \text{ when } x > 0.]
\end{aligned}$$

□

Proof of Proposition 2 (Equivalent as a Composition). Let $\boldsymbol{g} \in \mathbb{R}^d$ be an arbitrary gradient. To show this proposition, it is enough to show $\mathbb{P}\left\{[\mathcal{M}_{B,\beta}]_i(\boldsymbol{g}) = 1\right\} = \mathbb{P}\left\{[\mathcal{M}_{B,\mathsf{flip}} \circ \mathcal{M}_{B,0}]_i(\boldsymbol{g}) = 1\right\}$ holds for any $i \in [d]$.

To see this,

$$\begin{aligned}
\mathbb{P}\left\{[\mathcal{M}_{B,\mathsf{flip}} \circ \mathcal{M}_{B,0}]_i(\boldsymbol{g}) = 1\right\} &= \mathbb{P}\left\{[\mathcal{M}_{B,0}]_i(\boldsymbol{g}) = 1 \,\&\, \mathcal{M}_{B,\mathsf{flip}}(1) = 1\right\} \\
&\quad + \mathbb{P}\left\{[\mathcal{M}_{B,0}]_i(\boldsymbol{g}) = -1 \,\&\, \mathcal{M}_{B,\mathsf{flip}}(-1) = -1\right\} \\
&= \frac{B + \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}}{2B} \frac{2B+\beta}{2(B+\beta)} + \frac{B - \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}}{2B} \frac{\beta}{2(B+\beta)} \\
&= \frac{B + \beta + \mathsf{clip}\left\{\boldsymbol{g}_i, B\right\}}{2(B+\beta)} \\
&= \mathbb{P}\left\{[\mathcal{M}_{B,\beta}]_i(\boldsymbol{g}) = 1\right\}.
\end{aligned}$$

$\square$

D.3 CONVERGENCE RESULTS

Proposition 3 (Bounded Random Variable Variance Bound)**.** *Given a random variable X and a clipping threshold $B > 0$, if $\mu = \mathbb{E}[X] \in [-B, B]$, then* $\mathsf{var}\left(\mathsf{clip}\left(X, B\right)\right) \le \mathsf{var}\left(X\right) = \sigma^2$.

Proof of Proposition 3.

$$\begin{aligned}
\mathsf{var}\left(\mathsf{clip}\left(X, B\right)\right) :=& \mathbb{E}\left[\left(\mathsf{clip}(X,B) - \mathbb{E}\left[\mathsf{clip}(X,B)\right]\right)^2\right] \\
=& \mathbb{E}\left[\left(\mathsf{clip}(X,B) - \mathbb{E}\left[X\right]\right)^2\right] - \left(\mathbb{E}\left[\mathsf{clip}(X,B) - X\right]\right)^2 \\
\le& \mathbb{E}\left[\left(\mathsf{clip}(X,B) - \mathbb{E}\left[X\right]\right)^2\right]. \qquad (8)
\end{aligned}$$

For ease of exposition, we assume X admits a probability density function $f(x)$. General distributions of X can be shown analogously. It follows that

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathsf{clip}(X,B) - \mathbb{E}\left[X\right]\right)^2\right] \\
&= \int_B^\infty (B-\mu)^2 f(x)\mathrm{d}x + \int_{-B}^{B} (x-\mu)^2 f(x)\mathrm{d}x + \int_{-\infty}^{-B} (-B-\mu)^2 f(x)\mathrm{d}x \\
&\le \int_B^\infty (x-\mu)^2 f(x)\mathrm{d}x + \int_{-B}^{B} (x-\mu)^2 f(x)\mathrm{d}x + \int_{-\infty}^{-B} (x-\mu)^2 f(x)\mathrm{d}x \\
&= \mathsf{var}\left(X\right) = \sigma^2. \qquad (9)
\end{aligned}$$

Combining (8) and (9), we conclude $\mathsf{var}\left(\mathsf{clip}\left(X, B\right)\right) \le \mathsf{var}\left(X\right) = \sigma^2$. $\square$

D.3.1 SUB-GAUSSIAN AND HEAVY-TAILED DISTRIBUTIONS

Proof of Theorem 3 (Light and Heavy-tailed Sign Error). Recall that

$$\widehat{\boldsymbol{g}}_{mi}(t) = \begin{cases} [\mathcal{M}_{B,\beta}]_i\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j(t)\right) & \text{if } m \in \mathcal{N}(t); \\ * & \text{if } m \in \mathcal{B}(t), \end{cases}$$

where $*$ is an arbitrary value in {-1,1}. For any client $m \in [M]$ and any coordinate $i \in [d]$, let

$$X_{mi} = \mathbf{1}_{\{m \in \mathcal{S}(t)\}} \mathbf{1}_{\left\{\widehat{\boldsymbol{g}}_{mi} \neq \mathrm{sign}\left(\frac{1}{M}\sum_{m=1}^M \boldsymbol{g}_{mi}\right)\right\}},$$

$$\text{and} \qquad \widetilde{X}_{mi} = \mathbf{1}_{\{m \in \mathcal{S}(t)\}} \mathbf{1}_{\left\{[\mathcal{M}_\beta]_i\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j(t)\right) \neq \mathrm{sign}\left(\frac{1}{M}\sum_{m=1}^M \boldsymbol{g}_{mi}\right)\right\}}.$$

Notably, if $m \in \mathcal{B}(t)$, then it is possible that $X_{mi} \neq \widetilde{X}_{mi}$; otherwise, $X_{mi} = \widetilde{X}_{mi}$.

Without loss of generality, we assume the true aggregation is negative, i.e., $\mathrm{sign}\left(\nabla_i F\left(w(t)\right)\right) = -1$. The case when $\mathrm{sign}\left(\nabla_i F\left(w(t)\right)\right) = 1$ can be shown analogously.

For ease of exposition, we drop a condition of $w(t)$ in the conditional probability expressions unless otherwise noted. It holds that

$$\begin{aligned}
\mathbb{P}\left\{\text{sign}\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{g}}_{mi}\right)\neq -1\right\} &\leq \mathbb{P}\left\{\sum_{m=1}^{M}X_{mi}\geq\frac{|\mathcal{S}(t)|}{2}\right\}\\
&= \mathbb{P}\left\{\sum_{m\in\mathcal{N}(t)}\widetilde{X}_{mi}+\sum_{m\in\mathcal{B}(t)}X_{mi}\geq\frac{|\mathcal{S}(t)|}{2}\right\}\\
&= \mathbb{P}\left\{\sum_{m\in\mathcal{N}(t)}\widetilde{X}_{mi}\geq\frac{|\mathcal{S}(t)|}{2}-\sum_{m\in\mathcal{B}(t)}X_{mi}\right\}\\
&\leq \mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi}\geq\frac{|\mathcal{S}(t)|}{2}-\sum_{m\in\mathcal{B}(t)}X_{mi}\right\}. \qquad (10)
\end{aligned}$$

Next, we bound $\sum_{m=1}^{M}\widetilde{X}_{mi}$ and $\sum_{m\in\mathcal{B}(t)}X_{mi}$ separately.

When the system adversary is static, i.e., the system adversary does not know $\mathcal{S}(t)$, it corrupts clients independently of $\mathcal{S}(t)$. Hence,

$$\sum_{m\in\mathcal{B}(t)}X_{mi}\leq\sum_{m\in\mathcal{B}(t)}\mathbf{1}_{\{m\in\mathcal{S}(t)\}}. \qquad (11)$$

We know that if $\tau(t)\leq\frac{2}{p^2}\log\frac{6}{c}$, then $\sum_{m\in\mathcal{B}(t)}\mathbf{1}_{\{m\in\mathcal{S}(t)\}}\leq\frac{2}{p^2}\log\frac{6}{c}$. Otherwise, with probability at least $1-\frac{c}{6}$, it is true that $\sum_{m\in\mathcal{B}(t)}\mathbf{1}_{\{m\in\mathcal{S}(t)\}}\leq\frac{3}{2}p\tau(t)$.

On the other hand, when the system adversary is adaptive, it chooses $\mathcal{B}(t)$ based on $\mathcal{S}(t)$. In particular, if $|\mathcal{S}(t)|\leq\tau(t)$, then the adversary chooses $\mathcal{B}(t)=\mathcal{S}(t)$. Otherwise, i.e., $|\mathcal{S}(t)|>\tau(t)$, the adversary chooses an arbitrary subset of $\mathcal{S}(t)$. In both cases, it holds that

$$\sum_{m\in\mathcal{B}(t)}X_{mi}\leq\sum_{m\in\mathcal{B}(t)}\mathbf{1}_{\{m\in\mathcal{S}(t)\}}\leq\min\{\tau(t),|\mathcal{S}(t)|\}\leq\tau(t). \qquad (12)$$

For ease of exposition, we first focus on adaptive adversary and will visit the static adversary towards the end of this proof. Observe that $|\mathcal{S}(t)|=\sum_{m=1}^{M}\mathbf{1}_{\{m\in\mathcal{S}(t)\}}$. Let $\widetilde{Y}_{mi}=\widetilde{X}_{mi}-\frac{\mathbf{1}_{\{m\in\mathcal{S}(t)\}}}{2}$. Conditioning on the mini-batch stochastic gradients $\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}$, we have

$$\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]=\mathbb{E}\left[\widetilde{X}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]-\frac{p}{2}=\frac{p}{2B+2\beta}\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right).$$

Taking expectation over $\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}$, we get

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]\right]&=\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]-p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right]+\frac{p\boldsymbol{g}_{mi}}{2B+2\beta}\\
&=\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]-p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right]+\frac{p\boldsymbol{g}_{mi}}{2B+2\beta}.
\end{aligned} \qquad (13)$$

It turns out that $\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]-p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right]$ is small:

$$\begin{aligned}
&\frac{1}{p}\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi}\mid\boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]-p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right]\\
&=\underbrace{\frac{B\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\geq B\right\}-B\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\leq -B\right\}}{2B+2\beta}}_{\text{(A)}}+\underbrace{\frac{\mathbb{E}\left[-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\mathbf{1}_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\right|\geq B\right\}}\right]}{2B+2\beta}}_{\text{(B)}}.
\end{aligned}$$

We bound (A) and (B) for sub-Gaussian and heavy-tailed noise separately.

First, for sub-Gaussian distributions with Assumption 4, we have

$$\begin{aligned}
(\mathrm{A}) \leq & \frac{B}{2B+2\beta} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right] \geq B - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right\} \\
\leq & \frac{B}{2B+2\beta} \exp\left(-\frac{n\left(B-\boldsymbol{g}_{mi}\right)^{2}}{2\sigma_{mi}^{2}}\right) \\
\leq & \frac{B}{2B+2\beta} \exp\left(-\frac{n\epsilon_0^2 B_0^2}{2\sigma_{mi}^{2}}\right) \\
\leq & \frac{1}{2} \exp\left(-\frac{n}{2}\right) \quad [\text{since } \epsilon_0 > \frac{\sigma}{B_0}],
\end{aligned}$$

and

$$\begin{aligned}
(\mathrm{B}) = & \frac{\mathbb{E}\left[-\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} \mathbf{1}_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right| \geq B\right\}}\right]}{2B+2\beta} \\
= & \frac{\int_{-\infty}^{-B} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} < t\right\} \mathrm{d}t - \int_{B}^{+\infty} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} > t\right\} \mathrm{d}t}{2B+2\beta} \\
\leq & \frac{\int_{-\infty}^{-B} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right] < t - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right\} \mathrm{d}t}{2B+2\beta} \\
\leq & \frac{\int_{-\infty}^{-B} \exp\left(-\frac{(t-\boldsymbol{g}_{mi})^2}{2\sigma_{mi}^2/n}\right) \mathrm{d}t}{2B+2\beta} \quad [\text{Mill's ratio Gordon (1941)}] \\
= & \frac{1}{2B+2\beta} \int_{-\infty}^{-B} \left[-\frac{2\sigma_{mi}^2/n}{2\left(t-\boldsymbol{g}_{mi}\right)}\right] \left[-\frac{2\left(t-\boldsymbol{g}_{mi}\right)}{2\sigma_{mi}^2/n}\right] \exp\left[-\frac{\left(t-\boldsymbol{g}_{mi}\right)^2}{2\sigma_{mi}^2/n}\right] \mathrm{d}t \\
\leq & \frac{\sigma_{mi}^2/n}{\left(2B+2\beta\right)\left(B+\boldsymbol{g}_{mi}\right)} \int_{-\infty}^{-B} \left[-\frac{2\left(t-\boldsymbol{g}_{mi}\right)}{2\sigma_{mi}^2/n}\right] \exp\left(-\frac{\left(t-\boldsymbol{g}_{mi}\right)^2}{2\sigma_{mi}^2/n}\right) \mathrm{d}t \\
\leq & \frac{\sigma_{mi}^2}{n\epsilon_0 B_0 (2B+2\beta)} \exp\left(-\frac{n\epsilon_0^2 B_0^2}{2\sigma_{mi}^2}\right) \\
\leq & \frac{\sigma_{mi}^2}{2n\epsilon_0^2 B_0^2} \exp\left(-\frac{n\epsilon_0^2 B_0^2}{2\sigma_{mi}^2}\right) \quad [\beta > 0 \text{ and } B := (1+\epsilon_0)B_0 > \epsilon_0 B_0] \\
\leq & \frac{1}{2n} \exp\left(-\frac{n}{2}\right),
\end{aligned}$$

where the last inequality follows from the choice of $\epsilon_0 > \frac{\sigma}{B_0}$. Combining the bounds of (A) and (B), we get $\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] - p\frac{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right] \leq p\exp\left(-\frac{n}{2}\right)$. Hence,

$$\mathbb{E}\left[\widetilde{Y}_{mi}\right] \leq p\exp\left(-\frac{n}{2}\right) + \frac{p\boldsymbol{g}_{mi}}{2B+2\beta}. \tag{14}$$

Second, for heavy-tailed distributions with Assumption 5, we have

$$\begin{aligned}
(\mathrm{A}) \leq & \frac{B}{2B+2\beta} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right] \geq B - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right\} \\
\leq & \frac{B}{2B+2\beta} \mathbb{P}\left\{\left|\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right|^{p'} \geq n^{p'} \left|B - \boldsymbol{g}_{mi}\right|^{p'}\right\} \\
\leq & \frac{B}{2B+2\beta} \frac{\mathbb{E}\left[\left|\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right|^{p'}\right]}{n^{p'} \left|B - \boldsymbol{g}_{mi}\right|^{p'}} \quad [\text{Markov's inequality}] \\
\leq & \underbrace{\frac{B\sum_{j=1}^{n} \mathbb{E}\left[\left|\boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\boldsymbol{g}_{mi}^{j}\right]\right|^{p'}\right] + B\left(\sum_{j=1}^{n} \mathbb{E}\left[\left|\boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\boldsymbol{g}_{mi}^{j}\right]\right|^{2}\right]\right)^{\frac{p'}{2}}}{(2B+2\beta) n^{p'} \left|B - \boldsymbol{g}_{mi}\right|^{p'}}}_{\text{Rosenthal-type inequality Merlevède \& Peligrad (2013)}} \\
\leq & \frac{1}{2} \frac{n M_{p'} + n^{\frac{p'}{2}} M_{p'}}{n^{p'} \left|B - \boldsymbol{g}_{mi}\right|^{p'}} \quad [M_2^{\frac{1}{2}} \leq M_{p'}^{\frac{1}{p'}} \text{ for } p' \geq 4] \\
\leq & \frac{M_{p'}}{n^{\frac{p'}{2}} \epsilon_0^{p'} B_0^{p'}} \leq \frac{1}{n^{\frac{p'}{2}}}
\end{aligned}$$

and

$$\begin{aligned}
(\mathrm{B}) = & \frac{\mathbb{E}\left[-\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} \mathbf{1}_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right| \geq B\right\}}\right]}{2B+2\beta} \\
= & \frac{\int_{-\infty}^{-B} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} < t\right\} \mathrm{d}t - \int_{B}^{+\infty} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} > t\right\} \mathrm{d}t}{2B+2\beta} \\
\leq & \frac{\int_{-\infty}^{-B} \mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j} - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right] < t - \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}\right]\right\} \mathrm{d}t}{2B+2\beta} \\
\leq & \frac{1}{2B+2\beta} \int_{-\infty}^{-B} \frac{2M_{p'}}{n^{\frac{p'}{2}} \left|t - \boldsymbol{g}_{mi}\right|^{p'}} \mathrm{d}t \quad [\text{similar argument as in } (\mathrm{A})] \\
\leq & \frac{1}{2B+2\beta} \frac{1}{\epsilon_0^{p'-1} B_0^{p'-1} (p'-1) n^{\frac{p'}{2}}} \leq \frac{1}{(p'-1) n^{\frac{p'}{2}}} \leq \frac{1}{n^{\frac{p'}{2}}},
\end{aligned}$$

where the last inequality follows from the choice of $\epsilon_0 > \frac{M_{p'}^{\frac{1}{p'}}}{B_0}$. Combining the bounds of (A) and (B), we get $\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] - p\frac{\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right] \leq \frac{2p}{n^{\frac{p'}{2}}}$. Hence,

$$\mathbb{E}\left[\widetilde{Y}_{mi}\right] \leq \frac{2p}{n^{\frac{p'}{2}}} + \frac{p\boldsymbol{g}_{mi}}{2B+2\beta}. \tag{15}$$

Let us consider two mutually complement events $\mathcal{E}_1$ and $\mathcal{E}_2$:

$$\mathcal{E}_1 := \left\{\frac{1}{2(B+\beta)} \sum_{m=1}^{M} \mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}, B\right) - \mathbb{E}\left[\frac{1}{2(B+\beta)} \sum_{m=1}^{M} \mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}, B\right)\right] \leq \frac{c_0}{4(B+\beta)}\sqrt{M}\right\},$$

$$\mathcal{E}_2 := \left\{\frac{1}{2(B+\beta)} \sum_{m=1}^{M} \mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}, B\right) - \mathbb{E}\left[\frac{1}{2(B+\beta)} \sum_{m=1}^{M} \mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{g}_{mi}^{j}, B\right)\right] > \frac{c_0}{4(B+\beta)}\sqrt{M}\right\}.$$

We have

$$\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi}\geq\frac{|\mathcal{S}(t)|}{2}-\tau(t)\right\}\leq\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{Y}_{mi}\geq-\tau(t)\ \mid\ \mathcal{E}_1\right\}+\mathbb{P}\left\{\mathcal{E}_2\right\}. \tag{16}$$

By Proposition 3, we know that

$$\mathsf{var}\left(\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right)\right)\leq\mathsf{var}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\right)\leq\frac{1}{n}\mathsf{var}\left(\boldsymbol{g}_{mi}^{1}\right)=\frac{1}{n}\sigma_{mi}^{2}\leq\frac{1}{n}\sigma^{2}.$$

In addition, $\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right)$ is bounded and thus sub-Gaussian. Hence, we have

$$\mathbb{P}\left\{\mathcal{E}_2\right\}\leq\exp\left(-\frac{\frac{c_0^2M}{4}}{\frac{2M\sigma^2}{n}}\right).$$

Since $c_0\geq\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}$, we have $\mathbb{P}\left\{\mathcal{E}_2\right\}\leq\frac{c}{6}$.

For the first term in the right-hand side of Eq. (16), we have

$$\begin{aligned}&\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{Y}_{mi}\geq-\tau(t)\ \mid\ \mathcal{E}_1\right\}\\&=\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{Y}_{mi}-\mathbb{E}\left[\sum_{m=1}^{M}\widetilde{Y}_{mi}\ \mid\ \boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]\geq\underbrace{-\tau(t)-\mathbb{E}\left[\sum_{m=1}^{M}\widetilde{Y}_{mi}\ \mid\ \boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]}_{\text{(C)}}\ \mid\ \mathcal{E}_1\right\}\end{aligned}$$

Recall that $\mathbb{E}\left[\widetilde{Y}_{mi}\ \mid\ \boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]=\frac{p}{2B+2\beta}\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right)$. We have

$$\begin{aligned}\text{(C)}\mid\mathcal{E}_1&=-\tau(t)-\frac{p}{2B+2\beta}\sum_{m=1}^{M}\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right)\mid\mathcal{E}_1\\&\geq-\tau(t)-\mathbb{E}\left[\frac{p}{2B+2\beta}\sum_{m=1}^{M}\mathsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j},B\right)\right]-\frac{pc_0}{4(B+\beta)}\sqrt{M}\\&=-\tau(t)-\sum_{m=1}^{M}\mathbb{E}\left[\widetilde{Y}_{mi}\right]-\frac{pc_0}{4(B+\beta)}\sqrt{M}\\&\begin{cases}\geq-\tau(t)-Mp\exp\left(-\frac{n}{2}\right)-\frac{pM}{2(B+\beta)}\nabla_iF(w(t))-\frac{pc_0}{4(B+\beta)}\sqrt{M} & \text{[Sub-Gaussian Noise]}\\\geq-\tau(t)-\frac{2Mp}{n^{\frac{p'}{2}}}-\frac{pM}{2(B+\beta)}\nabla_iF(w(t))-\frac{pc_0}{4(B+\beta)}\sqrt{M} & \text{[Heavy-tailed Noise]}\end{cases}\end{aligned}$$

Recall that $\nabla_iF(w(t))<0$. When $\frac{pM}{2(B+\beta)}\left|\nabla_iF(w(t))\right|\geq\tau(t)+Mp\exp\left(-\frac{n}{2}\right)+\frac{pc_0}{2(B+\beta)}\sqrt{M}$ (sub-Gaussian noise) or when $\frac{pM}{2(B+\beta)}\left|\nabla_iF(w(t))\right|\geq\tau(t)+\frac{2Mp}{n^{\frac{p'}{2}}}+\frac{pc_0}{2(B+\beta)}\sqrt{M}$ (heavy-tailed noise), we get

$$\begin{aligned}\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{Y}_{mi}\geq-\tau(t)\ \mid\ \mathcal{E}_1\right\}\leq&\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{Y}_{mi}-\mathbb{E}\left[\sum_{m=1}^{M}\widetilde{Y}_{mi}\ \mid\ \boldsymbol{g}_{mi}^{1},\cdots,\boldsymbol{g}_{mi}^{n}\right]\geq\frac{pc_0}{4(B+\beta)}\sqrt{M}\ \mid\ \mathcal{E}_1\right\}\\\leq&\exp\left(-\frac{p^2c_0^2}{8(B+\beta)^2}\right)\\\leq&\frac{3-5c}{6},\end{aligned}$$

where the last inequality holds because $c_0 \ge \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}$.

Therefore, for adaptive system adversary, choosing $c_0 = \max\left\{\sqrt{\frac{8\sigma^2}{n}\log\frac{6}{c}}, \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}\right\}$, we conclude that if $\frac{pM}{2(B+\beta)}\left|\nabla_i F(w(t))\right| \ge \tau(t) + Mp\exp\left(-\frac{n}{2}\right) + \frac{pc_0}{2(B+\beta)}\sqrt{M}$ (sub-Gaussian Noise) or if $\frac{pM}{2(B+\beta)}\left|\nabla_i F(w(t))\right| \ge \tau(t) + \frac{2Mp}{n^{\frac{p'}{2}}} + \frac{pc_0}{2(B+\beta)}\sqrt{M}$ (heavy-tailed noise), then

$$\mathbb{P}\left\{\text{sign}\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{g}}_{mi}\right) \neq \text{sign}\left(\nabla_i F(w(t))\right) \mid w(t)\right\} \le \frac{1-c}{2}.$$

Otherwise, $\mathbb{P}\left\{\text{sign}\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{g}}_{mi}\right) \neq \text{sign}\left(\nabla_i F(w(t))\right) \mid w(t)\right\} \le 1$.

It remains to show the case for static adversary. When $\tau(t) \le \frac{2}{p^2}\log\frac{6}{c}$, we bound Eq. (10) as

$$\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi} \ge \frac{|\mathcal{S}(t)|}{2} - \sum_{m\in\mathcal{B}(t)}X_{mi}\right\} \le \mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi} \ge \frac{|\mathcal{S}(t)|}{2} - \tau(t)\right\}.$$

When $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$, we bound Eq. (10) as

$$\mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi} \ge \frac{|\mathcal{S}(t)|}{2} - \sum_{m\in\mathcal{B}(t)}X_{mi}\right\} \le \mathbb{P}\left\{\sum_{m=1}^{M}\widetilde{X}_{mi} \ge \frac{|\mathcal{S}(t)|}{2} - \frac{3p}{2}\tau(t)\right\} + \frac{c}{6}.$$

The remaining proof follows the above argument for adaptive adversary. $\square$

Proof of Theorem 4 (Sub-Gaussian and Heavy-tailed Convergence Rate). By Assumption 2, we have

$$\begin{aligned}
F\left(w(t+1)\right) - F\left(w(t)\right) &\le \langle \nabla F(w(t)), w(t+1) - w(t)\rangle + \frac{L}{2}\|w(t+1) - w(t)\|^2 \\
&= -\eta\sum_{i=1}^{d}\left|\nabla F(w(t))_i\right|\mathbf{1}_{\{\widetilde{\boldsymbol{g}}_i = \text{sign}\left(\nabla_i F(w(t))\right)\}} \\
&\quad + \eta\sum_{i=1}^{d}\left|\nabla F(w(t))_i\right|\mathbf{1}_{\{\widetilde{\boldsymbol{g}}_i \neq \text{sign}\left(\nabla_i F(w(t))\right)\}} + \frac{Ld}{2}\eta^2 \\
&= -\eta\|\nabla F(w(t))\|_1 + 2\eta\sum_{i=1}^{d}\left|\nabla F(w(t))_i\right|\mathbf{1}_{\{\widetilde{\boldsymbol{g}}_i \neq \text{sign}\left(\nabla_i F(w(t))\right)\}} + \frac{Ld}{2}\eta^2,
\end{aligned}$$

where $\nabla F(w(t))_i$ is the i-th coordinate of $\nabla F(w(t))$. Then, by conditioning on parameter $w(t)$, we get

$$\begin{aligned}
&\mathbb{E}\left[F\left(w(t+1)\right) - F\left(w(t)\right)\big|w(t)\right] \\
&\le \mathbb{E}\left[-\eta\|\nabla F(w(t))\|_1 + 2\eta\sum_{i=1}^{d}\left|\nabla F(w(t))_i\right|\mathbf{1}_{\{\widetilde{\boldsymbol{g}}_i \neq \text{sign}(\nabla F(w(t))_i)\}} + \frac{Ld}{2}\eta^2\right] \\
&= -\eta\|\nabla F(w(t))\|_1 + \frac{Ld}{2}\eta^2 + 2\eta\sum_{i=1}^{d}\left|\nabla F(w(t))_i\right|\mathbb{P}\left\{\widetilde{\boldsymbol{g}}_i \neq \text{sign}\left(\nabla F(w(t))_i\right)\right\}.
\end{aligned}$$

Recall that $\Xi_1(n) = 2(B+\beta)\exp\left(-\frac{n}{2}\right)$, and $\Xi_2(n) = \frac{4(B+\beta)}{n^{\frac{p'}{2}}}$. Define

$$\begin{cases}
\mathrm{A}_1 = \left\{\left|\nabla_i F(w(t))\right| \ge \frac{2(B+\beta)}{pM}\tau(t) + \frac{c_0}{\sqrt{M}} + 2(B+\beta)\exp\left(-\frac{n}{2}\right)\right\}; \\
\mathrm{A}_2 = \left\{\left|\nabla_i F(w(t))\right| \ge \frac{2(B+\beta)}{pM}\tau(t) + \frac{c_0}{\sqrt{M}} + \frac{4(B+\beta)}{n^{\frac{p'}{2}}}\right\}.
\end{cases}$$

In the following proof, we denote $\mathrm{A} = \mathrm{A}_1$, $\Xi(n) = \Xi_1(n)$ for sub-Gaussian noise and $\mathrm{A} = \mathrm{A}_2$, $\Xi(n) = \Xi_2(n)$ for heavy-tailed noise.

We now have two cases:

First, when the system adversary is adaptive or the system adversary is static but with $\tau(t) \leq \frac{2}{p^2} \log \frac{6}{c}$, then

$$\begin{aligned}
&\mathbb{E}\left[F\left(w(t+1)\right) - F\left(w(t)\right) \big| w(t)\right] \\
&= -\eta \|\nabla F(w(t))\|_1 + \frac{Ld}{2}\eta^2 + 2\eta \sum_{i=1}^{d} |\nabla F(w(t))_i| \, \mathbb{P}\left\{\widetilde{\boldsymbol{g}}_i \neq \operatorname{sign}\left(\nabla_i F(w(t))\right)\right\} \mathbf{1}_{\{\mathrm{A}\}} \\
&\quad + 2\eta \sum_{i=1}^{d} |\nabla F(w(t))_i| \, \mathbb{P}\left\{\widetilde{\boldsymbol{g}}_i \neq \operatorname{sign}\left(\nabla_i F(w(t))\right)\right\} \mathbf{1}_{\{\mathrm{A}^{\complement}\}} \\
&\leq -\eta \|\nabla F(w(t))\|_1 + \frac{Ld}{2}\eta^2 \\
&\quad + 2\eta \sum_{i=1}^{d} |\nabla F(w(t))_i| \, \frac{1-c}{2} \mathbf{1}_{\{\mathrm{A}\}} \\
&\quad + 2\eta \sum_{i=1}^{d} \left[\frac{2(B+\beta)\tau(t)}{pM} + \frac{c_0}{\sqrt{M}} + \Xi(n)\right] \mathbf{1}_{\{\mathrm{A}^{\complement}\}} \\
&\leq -\eta c \|\nabla F(w(t))\|_1 + \frac{Ld}{2}\eta^2 + 2\eta d \frac{c_0}{\sqrt{M}} + 4\eta d \frac{(B+\beta)\tau(t)}{pM} + 2\eta d \Xi(n).
\end{aligned}$$

Therefore, by Assumption 1, we have

$$\begin{aligned}
F^* - F(w(0)) &\leq \mathbb{E}\left[F\left(w(T)\right) - F\left(w(0)\right)\right] \\
&\leq -\eta c \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(w(t))\|_1\right] + \frac{\eta^2 L d T}{2} + 2\eta d T \frac{c_0}{\sqrt{M}} + 2\eta d T \Xi(n) + 4\eta d \frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pM}.
\end{aligned}$$

Rearrange the inequality and plug in $\eta = \frac{1}{\sqrt{dT}}$, we get

$$\begin{aligned}
\eta c \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(w(t))\|_1\right] &\leq F(w(0)) - F^* + \frac{\eta^2 L d T}{2} + 2\eta d T \frac{c_0}{\sqrt{M}} + 2\eta d T \Xi(n) + 4\eta d \frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pM} \\
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(w(t))\|_1\right] &\leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 4d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pTM} + 2d\Xi(n)\right].
\end{aligned}$$

Second, when the system adversary is static with $\tau(t) > \frac{2}{p^2} \log \frac{6}{c}$, follow a similar proof as above, we get

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 6d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM} + 2d\Xi(n)\right].$$

□

D.3.2 GAUSSIAN DISTRIBUTION

Proof of Corollary 2 (Gaussian Tail Sign Errors). Most of the proofs are the same with Theorem 3. We start from Eq. 13.

It turns out that $\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] - p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right]$ is small:

$$\begin{aligned}\frac{1}{p}\mathbb{E}\left[\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] - p\frac{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}}{2B+2\beta}\right] = & \underbrace{\frac{\left(B-\boldsymbol{g}_{mi}\right)\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j} \geq B\right\}}{2B+2\beta}}_{\text{(A)}} \\ & - \underbrace{\frac{\left(B+\boldsymbol{g}_{mi}\right)\mathbb{P}\left\{\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j} \leq -B\right\}}{2B+2\beta}}_{\text{(B)}} \\ & + \underbrace{\frac{\mathbb{E}\left[\left(-\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j} + \boldsymbol{g}_{mi}\right)\mathbf{1}_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}\right| \geq B\right\}}\right]}{2B+2\beta}}_{\text{(C)}}.\end{aligned} \tag{17}$$

We have,

$$\begin{aligned}&(2B+2\beta)\,\text{(A)} \\ &\leq \left(B-\boldsymbol{g}_{mi}\right)\cdot\frac{\sigma_{mi}/\sqrt{n}}{B-\boldsymbol{g}_{mi}}\cdot\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{\left(B-\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right) = \frac{\sigma_{mi}/\sqrt{n}}{\sqrt{2\pi}}\exp\left(-\frac{\left(B-\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right); \\ &(2B+2\beta)\,\text{(B)} \\ &\geq \left(B+\boldsymbol{g}_{mi}\right)\cdot\frac{\frac{B+\boldsymbol{g}_{mi}}{\sigma_{mi}/\sqrt{n}}}{\left(\frac{B+\boldsymbol{g}_{mi}}{\sigma_{mi}/\sqrt{n}}\right)^{2}+1}\cdot\frac{1}{\sqrt{2\pi}}\cdot\exp\left(-\frac{\left(B+\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right) \\ &= \left[1-\frac{\left(\sigma_{mi}/\sqrt{n}\right)^{2}}{\left(B+\boldsymbol{g}_{mi}\right)^{2}+\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right]\frac{\sigma_{mi}/\sqrt{n}}{\sqrt{2\pi}}\exp\left(-\frac{\left(B+\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right); \\ &(2B+2\beta)\,\text{(C)} \\ &= -\int_{B}^{\infty}\frac{x-\boldsymbol{g}_{mi}}{\sqrt{2\pi}\sigma_{mi}/\sqrt{n}}\exp\left(-\frac{\left(x-\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right)\mathrm{d}x - \int_{-\infty}^{-B}\frac{x-\boldsymbol{g}_{mi}}{\sqrt{2\pi}\sigma_{mi}/\sqrt{n}}\exp\left(-\frac{\left(x-\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right)\mathrm{d}x; \\ &= \frac{\sigma_{mi}/\sqrt{n}}{\sqrt{2\pi}}\left[\exp\left(-\frac{\left(B+\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right) - \exp\left(-\frac{\left(B-\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right)\right],\end{aligned}$$

where (A) and (B) follow because of Mill's ratio Gordon (1941).

Combining (A), (B), and (C), we get

$$\begin{aligned}(17) \leq & \frac{p\left(\sigma_{mi}/\sqrt{n}\right)^{3}}{\sqrt{2\pi}\left(2B+2\beta\right)\left[\left(B+\boldsymbol{g}_{mi}\right)^{2}+\left(\sigma_{mi}/\sqrt{n}\right)^{2}\right]}\exp\left(-\frac{\left(B+\boldsymbol{g}_{mi}\right)^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right)+\frac{p\boldsymbol{g}_{mi}}{2B+2\beta} \\ \leq & \frac{p\left(\sigma_{mi}/\sqrt{n}\right)^{3}}{\sqrt{2\pi}\left(2B+2\beta\right)\left[\epsilon_{0}^{2}B_{0}^{2}+\left(\sigma_{mi}/\sqrt{n}\right)^{2}\right]}\exp\left(-\frac{\epsilon_{0}^{2}B_{0}^{2}}{2\left(\sigma_{mi}/\sqrt{n}\right)^{2}}\right)+\frac{p\boldsymbol{g}_{mi}}{2B+2\beta} \\ \leq & \frac{p}{4\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right)+\frac{p\boldsymbol{g}_{mi}}{2B+2\beta},\end{aligned}$$

where the last inequality follows because $\epsilon_{0} > \frac{\sigma}{B_{0}}$ and $B := B_{0}+\epsilon_{0}B_{0} > \epsilon_{0}B_{0}$.

For the first term in the right hand side of Eq. (16), we have

$$
\begin{aligned}
&\mathbb{P}\left\{\sum_{m=1}^{M} \widetilde{Y}_{mi} \geq -\tau(t) \mid \mathcal{E}_1\right\} \\
=&\mathbb{P}\left\{\sum_{m=1}^{M} \widetilde{Y}_{mi} - \mathbb{E}\left[\sum_{m=1}^{M} \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] \geq \underbrace{-\tau(t) - \mathbb{E}\left[\sum_{m=1}^{M} \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right]}_{\text{(D)}} \mid \mathcal{E}_1\right\}
\end{aligned}
$$

Recall that $\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] = \frac{p}{2B+2\beta}\textsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}, B\right)$. We have

$$
\begin{aligned}
\text{(D)} \mid \mathcal{E}_1 &= -\tau(t) - \frac{p}{2B+2\beta}\sum_{m=1}^{M}\textsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}, B\right) \mid \mathcal{E}_1 \\
&\geq -\tau(t) - \mathbb{E}\left[\frac{p}{2B+2\beta}\sum_{m=1}^{M}\textsf{clip}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{g}_{mi}^{j}, B\right)\right] - \frac{pc_0}{4(B+\beta)}\sqrt{M} \\
&= -\tau(t) - \sum_{m=1}^{M}\mathbb{E}\left[\widetilde{Y}_{mi}\right] - \frac{pc_0}{4(B+\beta)}\sqrt{M} \\
&\geq -\tau(t) - \frac{Mp}{4\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) - \frac{p}{2(B+\beta)}\sum_{m=1}^{M}\boldsymbol{g}_{mi} - \frac{pc_0}{4(B+\beta)}\sqrt{M}
\end{aligned}
$$

Recall that $\nabla_i F(w(t)) < 0$. When $\frac{Mp}{2(B+\beta)}\left|\nabla_i F(w(t))\right| \geq \tau(t) + \frac{Mp}{4\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) + \frac{pc_0}{2(B+\beta)}\sqrt{M}$, we get

$$
\begin{aligned}
\mathbb{P}\left\{\sum_{m=1}^{M} \widetilde{Y}_{mi} \geq -\tau(t) \mid \mathcal{E}_1\right\} \leq&\mathbb{P}\left\{\sum_{m=1}^{M} \widetilde{Y}_{mi} - \mathbb{E}\left[\sum_{m=1}^{M} \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^{1}, \cdots, \boldsymbol{g}_{mi}^{n}\right] \geq \frac{pc_0}{4(B+\beta)}\sqrt{M} \mid \mathcal{E}_1\right\} \\
\leq&\exp\left(-\frac{p^2c_0^2}{8(B+\beta)^2}\right) \\
\leq&\frac{3-5c}{6},
\end{aligned}
$$

where the last inequality holds because $c_0 \geq \sqrt{\frac{8(B+\beta)^2}{p^2}\log\frac{6}{3-5c}}$.

The remaining proof follows the arguments in Theorem 3. □

Proof of Corollary 3 (Gaussian Tail Convergence Rate). This proof follows from Theorem 4. We also consider two cases here.

First, when the system adversary is adaptive or the system adversary is static but with $\tau(t) \leq \frac{2}{p^2}\log\frac{6}{c}$, plug in $\left|\nabla_i F(w(t))\right| \geq \frac{2(B+\beta)}{Mp}\tau(t) + \frac{(B+\beta)}{2\sqrt{2\pi}}\exp\left(-\frac{n}{2}\right) + \frac{c_0}{\sqrt{M}}$, we get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\Bigg[&\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + \frac{d}{\sqrt{2\pi}}(B+\beta)\exp\left(-\frac{n}{2}\right) \\
&+4d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pTM}\Bigg].
\end{aligned}
$$

Second, when the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$, plug in $|\nabla_i F(w(t))| \geq \frac{3(B+\beta)\tau(t)}{M} + \frac{(B+\beta)}{2\sqrt{2\pi}}\exp(-n/2) + \frac{c_0}{\sqrt{M}}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{(F(w(0))-F^*)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + \frac{d}{\sqrt{2\pi}}(B+\beta)\exp\left(-\frac{n}{2}\right) \right.$$
$$\left. +6d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM}\right].$$

□

D.4 BOUNDED STOCHASTIC GRADIENTS

Proof of Corollary 4 (Bounded Gradient Sign Errors). This proof follows from Theorem 3. Notably, if we choose $B = \widetilde{B}$, clip $\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j, B\right) = \frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j$ by Assumption 6. Thus, the bias introduced by the tail bound will be gone.

For the first term in the right-hand side of Eq. (16), we have

$$\begin{aligned}
&\mathbb{P}\left\{\sum_{m=1}^M \widetilde{Y}_{mi} \geq -\tau(t) \mid \mathcal{E}_1\right\} \\
=&\mathbb{P}\left\{\sum_{m=1}^M \widetilde{Y}_{mi} - \mathbb{E}\left[\sum_{m=1}^M \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^1,\cdots,\boldsymbol{g}_{mi}^n\right] \geq \underbrace{-\tau(t) - \mathbb{E}\left[\sum_{m=1}^M \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^1,\cdots,\boldsymbol{g}_{mi}^n\right]}_{\text{(A)}} \mid \mathcal{E}_1\right\}
\end{aligned}$$

Recall that $\mathbb{E}\left[\widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^1,\cdots,\boldsymbol{g}_{mi}^n\right] = \frac{p}{2B+2\beta}\frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j$. We have

$$\begin{aligned}
\text{(A)} \mid \mathcal{E}_1 &= -\tau(t) - \frac{p}{2B+2\beta}\sum_{m=1}^M \frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j \mid \mathcal{E}_1 \\
&\geq -\tau(t) - \mathbb{E}\left[\frac{p}{2B+2\beta}\sum_{m=1}^M \frac{1}{n}\sum_{j=1}^n \boldsymbol{g}_{mi}^j\right] - \frac{pc_0}{4(B+\beta)}\sqrt{M} \\
&= -\tau(t) - \sum_{m=1}^M \mathbb{E}\left[\widetilde{Y}_{mi}\right] - \frac{pc_0}{4(B+\beta)}\sqrt{M} \\
&\geq -\tau(t) - \frac{p}{2(B+\beta)}\sum_{m=1}^M \boldsymbol{g}_{mi} - \frac{pc_0}{4(B+\beta)}\sqrt{M}
\end{aligned}$$

Recall that $\nabla_i F(w(t)) < 0$. When $|\nabla_i F(w(t))| \geq \frac{2(B+\beta)\tau(t)}{Mp} + \frac{c_0}{\sqrt{M}}$, we get

$$\begin{aligned}
\mathbb{P}\left\{\sum_{m=1}^M \widetilde{Y}_{mi} \geq -\tau(t) \mid \mathcal{E}_1\right\} \leq& \mathbb{P}\left\{\sum_{m=1}^M \widetilde{Y}_{mi} - \mathbb{E}\left[\sum_{m=1}^M \widetilde{Y}_{mi} \mid \boldsymbol{g}_{mi}^1,\cdots,\boldsymbol{g}_{mi}^n\right] \geq \frac{pc_0}{4(B+\beta)}\sqrt{M} \mid \mathcal{E}_1\right\} \\
\leq& \exp\left(-\frac{p^2c_0^2}{8(B+\beta)^2}\right) \\
\leq& \frac{3-5c}{6},
\end{aligned}$$

The remaining proof also follows the arguments in Theorem 3. □

Proof of Corollary 5 (Bounded Gradient Convergence Rate). This proof follows from Theorem 4. We also consider two cases here.

First, when the system adversary is adaptive or the system adversary is static but with $\tau(t) \leq \frac{2}{p^2}\log\frac{6}{c}$, plug in $|F_i(w(t))| \geq \frac{2(B+\beta)\tau(t)}{Mp} + \frac{c_0}{\sqrt{M}}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 4d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{pTM}\right].$$

Second, when the system adversary is static with $\tau(t) > \frac{2}{p^2}\log\frac{6}{c}$, plug in $|\nabla_i F(w(t))| \geq \frac{3(B+\beta)\tau(t)}{M} + \frac{c_0}{\sqrt{M}}$, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(w(t))\|_1\right] \leq \frac{1}{c}\left[\frac{\left(F(w(0)) - F^*\right)\sqrt{d}}{\sqrt{T}} + \frac{L\sqrt{d}}{2\sqrt{T}} + 2d\frac{c_0}{\sqrt{M}} + 6d\frac{(B+\beta)\sum_{t=0}^{T-1}\tau(t)}{TM}\right].$$

□

E EXPERIMENT DETAILS

E.1 DATASETS AND PREPROCESSING

- **MNIST LeCun et al. (2009).** MNIST contains $60,000$ training images and $10,000$ testing images of 10 classes.
- **CIFAR-10 Krizhevsky et al. (2009).** CIFAR-10 contains $50,000$ training images and $10,000$ testing images of 10 classes.

Implementation. We build our codes upon PyTorch Paszke et al. (2019). We run all the experiments with 4 GPUs of type Tesla P100 and 1 GPU of type RTX 3060.

E.2 PARAMETERS

Communication rounds: 500 for both datasets in the section of client sampling. For the other sections, 80 and 300 communication rounds for MNIST and CIFAR-10, respectively.

Dataset partition: Clients' local datasets are evenly partitioned into balanced subsets. However, the distributions are non-IID since they follow Dirichlet distribution with a concentration α.

We consider a constant learning rate in all cases, and the choices are tuned through grid search. Specifically, $\eta \in \{0.0001, 0.001, 0.01, 0.1\}$, $B \in \{0.001, 0.01, 0.1, 1\}$. Although our theory indicates the algorithm is not sensitive to mini-batch size, we set a large batch size $n = 256$ for both datasets.

	Universal			β-Stochastic Sign SGD	FedAvg
	Learning Rate η	Mini-batch Size n	Hidden Units	B	Local Epochs
MNIST	0.01	256	64	0.01	1
CIFAR-10	0.01	256	200	0.01	1

Table 4: Hyperparameters

E.3 BYZANTINE BASELINE COMPARISONS

In this section, we reuse the network model and parameter settings in Table 4, set local epoch to be 1 for non-signed aggregation rules. We evaluate the algorithms on MNIST dataset and partition in a same manner as Appendix.E.1. We consider a total of 100 clients under full-participation with 20 Byzantine clients. The aggregation-rule-specific parameters are illustrated in the following part. All the experiment results are collected with 5 repetitions.

We compare our β stochastic sign compressor with Krum Blanchard et al. (2017), geometric median Chen et al. (2017), centered clipping Karimireddy et al. (2021) under three adversary models, including label flipping, inner product manipulation Xie et al. (2020), the "A little is enough" Baruch et al. (2019). Following Karimireddy et al. (2021), τ is set to be 10 in centered clipping since momentum is switched off. We first illustrate the adversary models below:

- **Label flipping**: Suppose original label is x, the adversary will replace it with $9 - x$;
- **Inner Product Manipulation**: The adversaries send $-\frac{\gamma}{|\mathcal{N}|}\sum_{i\in\mathcal{N}}\nabla f(\boldsymbol{w}_i)$, instead of honest messages, to mislead the parameter server, where ϵ is the strength of the adversary. Let $\gamma = 0.1$.
- **A Little is Enough**: The adversaries estimate the benign clients' mean $\mu_{\mathcal{N}}$ and standard deviation $\sigma_{\mathcal{N}}$. Then, they will construct new messages as $\mu_{\mathcal{N}} + z\sigma_{\mathcal{N}}$ and upload to the parameter server, where z is the strength of the adversary. We choose z according to Baruch et al. (2019):

$$z = \max_{z}\left(\Phi(z) < \frac{M-s}{M}\right),$$

where $z = \lfloor \frac{M}{2} + 1 \rfloor - |\mathcal{B}(t)|$, and Φ is the cumulative distribution function of standard normal distribution. For us, $z \approx 0.5$.

In Section 7, we present the performance of our compressor under flipping sign attacks. For sign-bit messages, this is the worst-case scenario as adversaries' messages cannot escape a binary value. Otherwise, it will be detected by PS and filtered out.

We consider a milder condition than the sign flipping attacks for a fair of competition. We allow adversaries to manipulate the mini-batch stochastic gradient but assume an honest compressor that will send out the correctly compressed corrupted messages to PS.

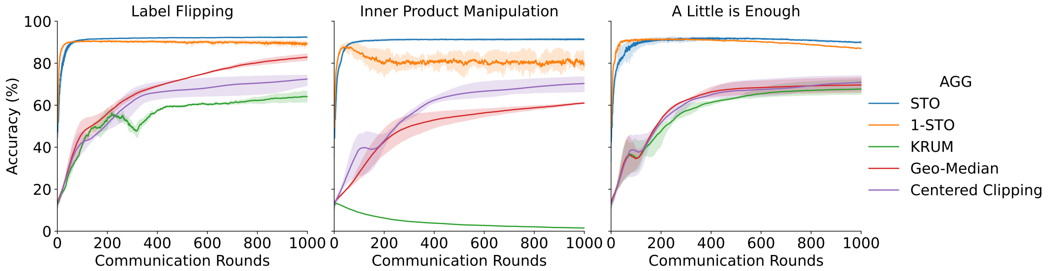

Figure 4: Comparisons with baselines: Krum, Geometric Median, Centered Clipping under Label Flipping, Inner Product Manipulation, and the ”A Little is Enough” Adversaries,
where 1-STO refers to our β stochastic sign compressor with $\beta = B = 0.01$.

Throughout the experiments, it is observed that our β stochastic sign compressor outperforms all other baseline algorithms when $\beta = 0$ or $\beta = B$. Notably, our compressor saves up to 31x communications and is differentially private when $\beta > 0$.