Supplementary Material for "Trust-Aware Topology Learning for Dynamic Decentralized Federated Learning under Adversaries"

Shubham Vaishnav, Murtaza Rangwala, Ali Beikmohammadi, Sindri Magnússon, and Rajkumar Buyya

This supplementary material provides: (i) statements and proofs of the three lemmas (structural invariants, consensus contraction, descent) supporting the convergence analysis of Section VI in the main paper; (ii) full proofs of Theorem 1, Corollary 1, and Proposition 1 stated in Section VI; (iii) the statement and proof of Theorem 2, extending Theorem 1 to $E \geq 1$ local SGD steps (referenced from but not stated in the main paper); and (iv) a distributed-backend verification confirming that the coordinator-free ZeroMQ implementation reproduces the simulation-backend accuracy trajectories reported in Section VII of the main paper.

References to assumptions, equations, and statements (e.g., Assumption 3, Theorem 1) refer to the main paper. Equation labels prefixed "S." are local to this document. We use $J := \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ for the averaging matrix, $U^t := (I-J)\Theta^t$ for the consensus error, G^t for the stacked stochastic gradients, and $f^\star := \inf_x f(x)$.

Notation Note: Consistent with Section VI of the main paper, when analyzing convergence under perfect screening, N denotes the size of the isolated compatible component $\mathcal{C}$, and $f(x)$ denotes the objective restricted to that component.

I. Structural properties of W^t

Lemma 1 (Structural invariants)**.** *For every epoch t, the matrix W^t produced by Algorithm 1 satisfies:*

1) Screened support: $w_{ij}^t = 0$ *for every* $j \notin C_i^t \cup \{i\}$.
2) Row-stochasticity: $w_{ij}^t \geq 0$ *and* $\sum_j w_{ij}^t = 1$ *for every* i.
3) Double stochasticity (under Assumption 3): $\sum_i w_{ij}^t = 1$ *for every* j.

The proof is given in the supplementary material.

To state the next results, we stack the local iterates and stochastic gradients into matrices

$$\Theta^t = \left[\theta_1^t, \dots, \theta_N^t\right]^\top, \quad G^t = \left[g_1(\theta_1^t, \xi_1^t), \dots, g_N(\theta_N^t, \xi_N^t)\right]^\top,$$

both in $\mathbb{R}^{N\times d}$, and define the *consensus error*

$$U^t = (I-J)\Theta^t,$$

which measures how far the local models drift from their mean. With $\hat{g}^t := \frac{1}{N}\sum_i g_i(\theta_i^t, \xi_i^t)$, double stochasticity of W^t implies that the average iterate $\bar{\theta}^t$ evolves according to

$$\bar{\theta}^{t+1} = \bar{\theta}^t - \eta\hat{g}^t, \tag{1}$$

which is the device that lets the analysis "see" the global objective f through the local dynamics.

A. Proof of Lemma 1 (Structural invariants of W^t)

Proof. Item 1 (screened support) is immediate from the aggregation step of Algorithm 1 (main paper, line 34), which restricts the sum to $j \in C_i^t \cup \{i\}$. Item 2 (row-stochasticity) follows from the normalization $\sum_{j\in C_i^t\cup\{i\}} w_{ij}^t = 1$ used in the aggregation equation of Section III, together with the nonnegativity of trust-based weights. Item 3 (double stochasticity) is the refinement provided by Assumption 3, which can be enforced by Metropolis–Hastings reweighting over the screened support. □

II. Consensus contraction lemma

Lemma 2 (Consensus contraction)**.** *Under Assumptions 1–4, for every $t \geq 0$,*

$$\mathbb{E}\|U^{t+1}\|_F^2 \leq \left(1 - \tfrac{p}{2}\right)\mathbb{E}\|U^t\|_F^2 + \frac{2\eta^2}{p}\,\mathbb{E}\|(I-J)G^t\|_F^2. \tag{2}$$

The proof is given in the supplementary material.

A. Proof of Lemma 2 (Consensus contraction)

Proof. The local-update and aggregation steps yield $\Theta^{t+1} = W^t(\Theta^t - \eta G^t)$. By double stochasticity (Assumption 3), $W^tJ = JW^t = J$ and $J^2 = J$, so $(W^t - J)J = 0$. Subtracting the mean rows from Θ^t and G^t inside the product is therefore a no-op:

$$\begin{aligned} U^{t+1} &= (I-J)\,\Theta^{t+1} \\ &= (W^t - J)(\Theta^t - \eta G^t) \\ &= (W^t - J)\big(U^t - \eta(I-J)G^t\big). \end{aligned}$$

Let $V^t := (I-J)G^t$. Both U^t and V^t have zero column means, so by Assumption 4 (consensus contraction),

$$\|(W^t - J)(U^t - \eta V^t)\|_F^2 \leq (1-p)\,\|U^t - \eta V^t\|_F^2.$$

Apply Young's inequality with parameter $\alpha = \frac{p}{2(1-p)} > 0$ (the case $p = 1$ is trivial):

$$\|U^t - \eta V^t\|_F^2 \leq (1+\alpha)\,\|U^t\|_F^2 + \left(1 + \tfrac{1}{\alpha}\right)\eta^2\|V^t\|_F^2.$$

With this choice of α:

$$(1-p)(1+\alpha) = 1 - \tfrac{p}{2}, \quad (1-p)\left(1+\tfrac{1}{\alpha}\right) = \tfrac{(1-p)(2-p)}{p} \leq \tfrac{2}{p}.$$

Combining and taking expectation yields the bound of Lemma 2. □

III. Descent lemma on the average iterate

Lemma 3 (Descent inequality)**.** *Under Assumptions 1–3 and stepsize $\eta \le \frac{1}{L}$,*

$$\begin{aligned}\mathbb{E}[f(\bar\theta^{t+1})] \le \mathbb{E}[f(\bar\theta^{t})] - \frac{\eta}{2}\,\mathbb{E}\|\nabla f(\bar\theta^{t})\|^2 \\ + \frac{\eta L^2}{2N}\,\mathbb{E}\|U^t\|_F^2 + \frac{L\eta^2\sigma^2}{2N}.\end{aligned} \tag{3}$$

The proof is given in the supplementary material.

A. Proof of Lemma 3 (Descent inequality)

Proof. L-smoothness of f applied to the average-iterate update $\bar\theta^{t+1} = \bar\theta^{t} - \eta\hat g^t$ gives

$$f(\bar\theta^{t+1}) \le f(\bar\theta^{t}) - \eta\langle\nabla f(\bar\theta^{t}), \hat g^t\rangle + \frac{L\eta^2}{2}\|\hat g^t\|^2.$$

Define the local-mean gradient $h^t := \frac{1}{N}\sum_i \nabla F_i(\theta_i^t)$, so that $\mathbb{E}[\hat g^t \mid \mathcal{F}^t] = h^t$, where $\mathcal{F}^t$ is the σ-algebra of all randomness through round t. By Assumption 2 and independence of g_i across clients, $\mathrm{Var}(\hat g^t \mid \mathcal{F}^t) \le \sigma^2/N$, so

$$\mathbb{E}\big[\|\hat g^t\|^2 \mid \mathcal{F}^t\big] \le \|h^t\|^2 + \tfrac{\sigma^2}{N}.$$

Taking conditional expectation in the smoothness inequality:

$$\begin{aligned}\mathbb{E}[f(\bar\theta^{t+1}) \mid \mathcal{F}^t] \le f(\bar\theta^{t}) - \eta\langle\nabla f(\bar\theta^{t}), h^t\rangle \\ + \frac{L\eta^2}{2}\|h^t\|^2 + \frac{L\eta^2\sigma^2}{2N}.\end{aligned} \tag{4}$$

Apply the polarization identity $\langle a, b\rangle = \frac{1}{2}(\|a\|^2 + \|b\|^2 - \|a-b\|^2)$:

$$\begin{aligned}&-\eta\langle\nabla f(\bar\theta^{t}), h^t\rangle + \tfrac{L\eta^2}{2}\|h^t\|^2 \\ &\quad= -\tfrac{\eta}{2}\|\nabla f(\bar\theta^{t})\|^2 + \tfrac{\eta}{2}\|\nabla f(\bar\theta^{t}) - h^t\|^2 \\ &\qquad+ \big(\tfrac{L\eta^2}{2} - \tfrac{\eta}{2}\big)\|h^t\|^2.\end{aligned}$$

For $\eta \le 1/L$, the coefficient of $\|h^t\|^2$ is nonpositive, so that term may be dropped. Jensen's inequality and L-smoothness of each F_i bound the bias:

$$\begin{aligned}\|\nabla f(\bar\theta^{t}) - h^t\|^2 &= \Big\|\tfrac{1}{N}\sum_i\big[\nabla F_i(\bar\theta^{t}) - \nabla F_i(\theta_i^t)\big]\Big\|^2 \\ &\le \tfrac{L^2}{N}\,\|U^t\|_F^2.\end{aligned}$$

Substituting back into (4) and taking total expectation yields Lemma 3. □

IV. Proof of Theorem 1 (Convergence under perfect screening, $E = 1$)

Proof. We combine Lemmas 2 and 3 through a coupled-inequality argument. Throughout, define

$$\begin{aligned}&a_t := \mathbb{E}\|U^t\|_F^2, \quad b_t := \mathbb{E}\|(I-J)G^t\|_F^2, \\ &S := \sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar\theta^{t})\|^2.\end{aligned}$$

a) Step 1: Cumulative consensus error: Lemma 2 gives $a_{t+1} \le (1-\frac{p}{2})a_t + \frac{2\eta^2}{p}b_t$. With $a_0 = 0$ (consensus initialization), iterating yields $a_t \le \frac{2\eta^2}{p}\sum_{s=0}^{t-1}(1-\frac{p}{2})^{t-1-s}b_s$. Summing over t and exchanging summation order,

$$\sum_{t=0}^{T-1} a_t \le \frac{2\eta^2}{p}\sum_{s=0}^{T-1} b_s \sum_{k=0}^{\infty}\Big(1-\tfrac{p}{2}\Big)^k = \frac{4\eta^2}{p^2}\sum_{t=0}^{T-1} b_t. \tag{5}$$

b) Step 2: Bound on b_t: The variance identity $\sum_i\|g_i^t - \hat g^t\|^2 \le \sum_i\|g_i^t\|^2$ combined with Assumption 2 ($\mathbb{E}\|g_i^t - \nabla F_i(\theta_i^t)\|^2 \le \sigma^2$) gives

$$b_t \le \sum_{i=1}^{N}\mathbb{E}\|\nabla F_i(\theta_i^t)\|^2 + N\sigma^2.$$

Apply $\|a\|^2 \le 2\|a-b\|^2 + 2\|b\|^2$ twice (first with $b = \nabla F_i(\bar\theta^t)$, then with $b = \nabla f(\bar\theta^t)$), use L-smoothness, sum over i, and apply Assumption 5:

$$\sum_i\|\nabla F_i(\theta_i^t)\|^2 \le 2L^2\|U^t\|_F^2 + 4N\tau_{\mathrm{scr}}^2 + 4N\|\nabla f(\bar\theta^t)\|^2.$$

Taking expectation,

$$\begin{aligned}b_t \le 2L^2 a_t + 4N\tau_{\mathrm{scr}}^2 + N\sigma^2 \\ + 4N\,\mathbb{E}\|\nabla f(\bar\theta^t)\|^2.\end{aligned} \tag{6}$$

c) Step 3: Closing the recursion: Substituting (6) into (5) and collecting the $\sum_t a_t$ term on the left:

$$\Big(1 - \tfrac{8\eta^2L^2}{p^2}\Big)\sum_t a_t \le \tfrac{4\eta^2N}{p^2}\big[4T\tau_{\mathrm{scr}}^2 + T\sigma^2 + 4S\big].$$

For $\eta \le p/(4L)$, the left factor is at least $1/2$, so

$$\sum_t a_t \le \frac{32N\eta^2}{p^2}\big[T(\tau_{\mathrm{scr}}^2 + \sigma^2) + S\big]. \tag{7}$$

d) Step 4: Combining with the descent lemma: Summing Lemma 3 over $t = 0, \dots, T-1$ and telescoping the f-difference:

$$\tfrac{\eta}{2}\,S \le f(\bar\theta^0) - f^\star + \tfrac{\eta L^2}{2N}\sum_t a_t + \tfrac{L\eta^2\sigma^2T}{2N}.$$

Substituting (7),

$$\begin{aligned}\tfrac{\eta}{2}\,S \le (f^0 - f^\star) + \tfrac{L\eta^2\sigma^2T}{2N} \\ + \tfrac{16L^2\eta^3}{p^2}\big[T(\tau_{\mathrm{scr}}^2 + \sigma^2) + S\big].\end{aligned}$$

For $\eta \le p/(8L)$, $\frac{16L^2\eta^3}{p^2} \le \frac{\eta}{4}$, so the S-term on the right is absorbed:

$$\begin{aligned}\tfrac{\eta}{4}\,S \le (f^0 - f^\star) + \tfrac{L\eta^2\sigma^2T}{2N} \\ + \tfrac{16L^2\eta^3T(\tau_{\mathrm{scr}}^2+\sigma^2)}{p^2}.\end{aligned}$$

Dividing both sides by $\eta T/4$ yields the bound stated in Theorem 1. □

V. Extension to local SGD ($E \geq 1$)

The convergence result extends to the case where each client performs $E \geq 1$ local SGD steps per round (as in Algorithm 1, line 24), at the cost of one additional assumption: the second moment of the stochastic gradient must be uniformly bounded (Assumption 6). This is a stronger condition than Assumption 2 (bounded *variance*), and it is what controls the local drift $\|\theta_i^{t,k} - \theta_i^t\|$ over the E local steps.

The local update is parameterized by a local stepsize η_l, with $\theta_i^{t,0} = \theta_i^t$ and

$$\theta_i^{t,k+1} = \theta_i^{t,k} - \eta_l\, g_i(\theta_i^{t,k}, \xi_i^{t,k}), \quad k = 0, \dots, E-1,$$

followed by aggregation $\theta_i^{t+1} = \sum_j w_{ij}^t \theta_j^{t,E}$. Define the effective per-round stepsize $\eta := E\eta_l$.

a) Assumption 6: Bounded gradient second moments.: There exists $G \geq 0$ such that for all $i \in \mathcal{V}$, $t \geq 0$, and points x visited by the iterates,

$$\mathbb{E}\|g_i(x, \xi_i)\|^2 \leq G^2.$$

This stronger assumption is invoked only in Theorem 2 (local SGD with $E > 1$) and Proposition 1 (imperfect screening). It is not needed for Theorem 1. Note that $G^2 \geq \sigma^2 + \sup_x \|\nabla f(x)\|^2$, and the boundedness typically follows from a uniform Lipschitz/bounded-gradient regularization of the loss.

Theorem 2 (Convergence with E local steps, perfect screening)**.** *Suppose Assumptions 1–6 hold, screening is perfect ($\delta_{\max} = 0$, $\Delta_{\text{scr}}^2 = 0$), and the iterates are initialized at consensus. If the local stepsize satisfies $\eta_l \leq \frac{1}{2LE}$ (equivalently, $\eta \leq \frac{1}{2L}$), then for any horizon $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big\|\nabla f(\bar\theta^t)\big\|^2 \leq \frac{2\big(f(\bar\theta^0) - f^\star\big)}{\eta T} + \frac{2L\eta\,\sigma^2}{NE} + \frac{(8 + \frac{2}{3}p^2)\,L^2\,\eta^2\,G^2}{p^2}. \tag{8}$$

The proof is given in the supplementary material.

Remark 1 (Effect of $E > 1$ and scope). Relative to Theorem 1, two effects are visible in (8): (i) the noise floor scales as $\sigma^2/(NE)$ rather than σ^2/N, reflecting the variance reduction from E stochastic gradients per round; and (ii) the heterogeneity-and-drift term is now expressed via G^2, which under $G^2 \gtrsim \tau_{\text{scr}}^2 + \sigma^2$ is comparable to (but generally looser than) the corresponding term in Theorem 1. The bounded-gradient Assumption 6 is restrictive (it does not hold for unregularized least-squares or logistic regression on unbounded data); a more refined analysis along the lines of [1] recovers the same $\mathcal{O}(\sigma/\sqrt{NET} + 1/T)$ rate without bounded gradients, at the cost of more involved constants.

A. Proof of Theorem 2 (Convergence with $E \geq 1$ local steps)

Proof. For $i \in \mathcal{V}$ and $k = 0, \dots, E$, let $\theta_i^{t,k}$ denote the local iterate after k local steps in round t, with $\theta_i^{t,0} = \theta_i^t$. Stack the per-step stochastic gradients as $\widetilde{G}_i^t := \sum_{k=0}^{E-1} g_i(\theta_i^{t,k}, \xi_i^{t,k})$, so the aggregation step reads $\Theta^{t+1} = W^t(\Theta^t - \eta_l \widetilde{G}^t)$. Let $\hat g^t := \frac{1}{NE}\sum_{i,k} g_i(\theta_i^{t,k}, \xi_i^{t,k})$ be the all-sample average. Double stochasticity gives the average-iterate update $\bar\theta^{t+1} = \bar\theta^t - \eta \hat g^t$ with $\eta = E\eta_l$. Define the true local-gradient average as $\bar h^t := \frac{1}{NE}\sum_{i,k} \nabla F_i(\theta_i^{t,k})$.

a) Step 1: Variance of $\hat g^t$: Decompose the update into the true gradient and the stochastic noise:

$$\hat g^t = \bar h^t + \underbrace{\tfrac{1}{NE}\sum_{i,k}\big[\,g_i(\theta_i^{t,k}, \xi_i^{t,k}) - \nabla F_i(\theta_i^{t,k})\,\big]}_{=:\hat g^t_{\text{noise}}}.$$

Let $\mathcal{F}^{t,k}$ be the filtration capturing all randomness up to local step k in round t. The noise terms $g_i(\theta_i^{t,k}, \xi_i^{t,k}) - \nabla F_i(\theta_i^{t,k})$ are martingale differences with respect to $\mathcal{F}^{t,k}$ and are independent across clients. Thus, their sum has variance:

$$\mathbb{E}\|\hat g^t_{\text{noise}}\|^2 \leq \frac{NE\,\sigma^2}{(NE)^2} = \frac{\sigma^2}{NE}.$$

Because the trajectory $\theta_i^{t,k}$ depends on past noise, $\bar h^t$ and $\hat g^t_{\text{noise}}$ are not strictly orthogonal. Applying the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$:

$$\mathbb{E}\|\hat g^t\|^2 \leq 2\mathbb{E}\|\bar h^t\|^2 + \frac{2\sigma^2}{NE}. \tag{9}$$

b) Step 2: Local-drift bound (Assumption 6): For $k \in \{0, \dots, E-1\}$, the telescoping identity $\theta_i^{t,k} - \theta_i^t = -\eta_l \sum_{\ell=0}^{k-1} g_i(\theta_i^{t,\ell}, \xi_i^{t,\ell})$ combined with Cauchy–Schwarz and Assumption 6 yields

$$\mathbb{E}\|\theta_i^{t,k} - \theta_i^t\|^2 \leq \eta_l^2\, k \sum_{\ell=0}^{k-1} \mathbb{E}\|g_i^{t,\ell}\|^2 \leq \eta_l^2\, k^2\, G^2. \tag{10}$$

c) Step 3: Consensus contraction (local version): By the same algebra as Lemma 2 applied to $\Theta^{t+1} = W^t(\Theta^t - \eta_l \widetilde{G}^t)$:

$$\mathbb{E}\|U^{t+1}\|_F^2 \leq \Big(1 - \tfrac{p}{2}\Big)\mathbb{E}\|U^t\|_F^2 + \tfrac{2\eta_l^2}{p}\mathbb{E}\|(I-J)\widetilde{G}^t\|_F^2.$$

Bounding the second factor using $\|\widetilde{G}_i^t\|^2 \leq E\sum_k \|g_i^{t,k}\|^2$ (Cauchy–Schwarz) and Assumption 6 gives $\mathbb{E}\|(I-J)\widetilde{G}^t\|_F^2 \leq NE^2G^2$. Using $\eta = E\eta_l$:

$$\mathbb{E}\|U^{t+1}\|_F^2 \leq \Big(1 - \tfrac{p}{2}\Big)\mathbb{E}\|U^t\|_F^2 + \tfrac{2N\eta^2G^2}{p}.$$

Iterating from $U^0 = 0$ and summing the geometric series:

$$\sum_{t=0}^{T-1} \mathbb{E}\|U^t\|_F^2 \leq \frac{4TN\eta^2G^2}{p^2}. \tag{11}$$

d) Step 4: Descent inequality (local version): From L-smoothness on $\bar\theta^{t+1} = \bar\theta^t - \eta\hat g^t$ and taking total expectation:

$$\mathbb{E}[f(\bar\theta^{t+1})] \leq \mathbb{E}[f(\bar\theta^t)] - \eta\mathbb{E}\langle\nabla f(\bar\theta^t), \hat g^t\rangle + \frac{L\eta^2}{2}\mathbb{E}\|\hat g^t\|^2.$$

Since $\mathbb{E}[\hat g^t] = \mathbb{E}[\bar h^t]$, we apply (9) and the polarization identity $-\eta\langle a, b\rangle = -\frac{\eta}{2}\|a\|^2 - \frac{\eta}{2}\|b\|^2 + \frac{\eta}{2}\|a-b\|^2$:

$$\begin{aligned}
\mathbb{E}[f(\bar\theta^{t+1})] &\leq \mathbb{E}[f(\bar\theta^t)] - \eta\mathbb{E}\langle\nabla f(\bar\theta^t), \bar h^t\rangle + L\eta^2\mathbb{E}\|\bar h^t\|^2 \\
&\quad + \frac{L\eta^2\sigma^2}{NE} \\
&= \mathbb{E}[f(\bar\theta^t)] - \frac{\eta}{2}\mathbb{E}\|\nabla f(\bar\theta^t)\|^2 \\
&\quad + \frac{\eta}{2}\mathbb{E}\|\nabla f(\bar\theta^t) - \bar h^t\|^2 - \Big(\frac{\eta}{2} - L\eta^2\Big)\mathbb{E}\|\bar h^t\|^2 \\
&\quad + \frac{L\eta^2\sigma^2}{NE}.
\end{aligned}$$

For $\eta \le \frac{1}{2L}$, the term $(\frac{\eta}{2} - L\eta^2) \ge 0$, allowing us to drop the negative $\|\bar{h}^t\|^2$ term. Bounding the bias via Jensen, L-smoothness, and $\|\bar{\theta}^t - \theta_i^{t,k}\|^2 \le 2\|U_i^t\|^2 + 2\|\theta_i^t - \theta_i^{t,k}\|^2$:

$$\mathbb{E}\|\nabla f(\bar{\theta}^t) - \bar{h}^t\|^2 \le \frac{L^2}{NE}\sum_{i,k}\mathbb{E}\|\bar{\theta}^t - \theta_i^{t,k}\|^2 \le \frac{2L^2}{N}\mathbb{E}\|U^t\|_F^2 + \frac{2L^2}{NE}\sum_{i,k}\mathbb{E}\|\theta_i^t - \theta_i^{t,k}\|^2.$$

Apply (10) and $\sum_{k=0}^{E-1} k^2 \le E^3/3$:

$$\sum_{i,k}\mathbb{E}\|\theta_i^t - \theta_i^{t,k}\|^2 \le \frac{N\eta_l^2 G^2 E^3}{3}.$$

With $\eta_l^2 E^2 = \eta^2$, this gives $\frac{2L^2}{NE} \cdot \frac{N\eta_l^2 G^2 E^3}{3} = \frac{2L^2\eta^2 G^2}{3}$. Combining:

$$\mathbb{E}[f(\bar{\theta}^{t+1})] \le \mathbb{E}[f(\bar{\theta}^t)] - \frac{\eta}{2}\mathbb{E}\|\nabla f(\bar{\theta}^t)\|^2 + \frac{\eta L^2}{N}\mathbb{E}\|U^t\|_F^2 + \frac{\eta^3 L^2 G^2}{3} + \frac{L\eta^2\sigma^2}{NE}. \tag{12}$$

e) Step 5: Combine: Summing (12) over t and using (11):

$$\begin{aligned}\frac{\eta}{2}\sum_t \mathbb{E}\|\nabla f(\bar{\theta}^t)\|^2 &\le (f^0 - f^\star) + \frac{TL\eta^2\sigma^2}{NE} + \frac{\eta L^2}{N}\cdot\frac{4TN\eta^2G^2}{p^2} + \frac{T\eta^3L^2G^2}{3}\\ &= (f^0 - f^\star) + \frac{TL\eta^2\sigma^2}{NE} + T\eta^3L^2G^2\Big(\frac{4}{p^2} + \frac{1}{3}\Big).\end{aligned}$$

Dividing by $\frac{\eta T}{2}$ and noting that $\frac{8}{p^2} + \frac{2}{3} = \frac{8+2p^2/3}{p^2}$, we obtain the final bound:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big\|\nabla f(\bar{\theta}^t)\big\|^2 \le \frac{2\big(f(\bar{\theta}^0) - f^\star\big)}{\eta T} + \frac{2L\eta\,\sigma^2}{NE} + \frac{(8 + \frac{2}{3}p^2)\,L^2\,\eta^2\,G^2}{p^2}.$$

□

VI. Proof of Corollary 1 (Screening improves the bound)

Proof. The right-hand side of the bound in Theorem 1, when applied to a specific objective and communication graph, is monotone increasing in the data heterogeneity τ^2 and monotone decreasing in the consensus rate p. Substituting the stated inequalities $\tau_{\mathrm{scr}}^2 \le \tau_{\mathrm{raw}}^2$ and $p_{\mathrm{scr}} \ge p_{\mathrm{raw}} - \epsilon$ into the heterogeneity-and-topology term of Theorem 1 (applied with the screened component parameters τ_{scr}^2 and p_{scr}) directly yields the claim. When screening preserves component connectivity ($\epsilon = 0$), the bound strictly improves whenever component isolation successfully reduces data heterogeneity ($\tau_{\mathrm{scr}}^2 < \tau_{\mathrm{raw}}^2$). □

VII. Proof of Proposition 1 (Imperfect screening)

Proof. Let $\mathcal{C}$ be the target honest component of size N. For $i \in \mathcal{C}$, the actual aggregation step is $\theta_i^{t+1} = \sum_{j\in\mathcal{C}} w_{ij}^t\theta_j^{t+1/2} + \beta_{i,\mathrm{adv}}^t$, where the adversarial injection is $\beta_{i,\mathrm{adv}}^t = \sum_{b\in\mathcal{B}} w_{ib}^t\theta_b^{B,t+1/2}$. Define the matrix $B^t \in \mathbb{R}^{N\times d}$ whose i-th row is $\beta_{i,\mathrm{adv}}^t$. By Assumption 8, $\|B^t\|_F \le \sqrt{N}\delta_{\max}\Theta_B$.

We rewrite the global update over $\mathcal{C}$ by injecting and subtracting the ideal doubly stochastic matrix $\widetilde{W}^t$:

$$\Theta^{t+1} = \widetilde{W}^t(\Theta^t - \eta G^t) + \underbrace{(W^t - \widetilde{W}^t)(\Theta^t - \eta G^t) + B^t}_{=:E^t}.$$

a) Step 1: Bounding the perturbation E^t: Taking the Frobenius norm and applying the triangle inequality:

$$\|E^t\|_F \le \|W^t - \widetilde{W}^t\|_F\big(\|\Theta^t\|_F + \eta\|G^t\|_F\big) + \|B^t\|_F.$$

By Assumption 9 (bounded domain), $\|\Theta^t\|_F \le \sqrt{N}R$. By Assumption 6, the expected stochastic gradient norm is bounded by $\mathbb{E}[\|G^t\|_F] \le \sqrt{N}G$. Taking the expectation and applying Jensen's inequality to the screening error ($\mathbb{E}\|W^t - \widetilde{W}^t\|_F \le \Delta_{\mathrm{scr}}$):

$$\mathbb{E}\|E^t\|_F \le \sqrt{N}\big[\Delta_{\mathrm{scr}}(R + \eta G) + \delta_{\max}\Theta_B\big] = \sqrt{N}\rho. \tag{13}$$

b) Step 2: Average-iterate update and Descent Lemma: Multiplying the update by $\frac{1}{N}\mathbf{1}^\top$, and noting $\mathbf{1}^\top\widetilde{W}^t = \mathbf{1}^\top$, the average iterate evolves as:

$$\bar{\theta}^{t+1} = \bar{\theta}^t - \eta\hat{g}^t + e^t,$$

where $e^t = \frac{1}{N}\mathbf{1}^\top E^t$. By Jensen's inequality, $\mathbb{E}\|e^t\|^2 \le \frac{1}{N}\mathbb{E}\|E^t\|_F^2 \le \rho^2$. By the L-smoothness of f:

$$f(\bar{\theta}^{t+1}) \le f(\bar{\theta}^t) + \langle\nabla f(\bar{\theta}^t), -\eta\hat{g}^t + e^t\rangle + \frac{L}{2}\|-\eta\hat{g}^t + e^t\|^2.$$

Take the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}^t]$. Noting $\mathbb{E}[\hat{g}^t] = h^t$:

$$\mathbb{E}[f^{t+1}] \le f^t - \eta\langle\nabla f(\bar{\theta}^t), h^t\rangle + \mathbb{E}\langle\nabla f(\bar{\theta}^t), e^t\rangle + \tfrac{L}{2}\mathbb{E}\|{-\eta\hat{g}^t} + e^t\|^2.$$

We bound the cross terms using Young's inequality:

$$\langle\nabla f(\bar{\theta}^t), e^t\rangle \le \frac{\eta}{8}\|\nabla f(\bar{\theta}^t)\|^2 + \frac{2}{\eta}\|e^t\|^2$$
$$\|-\eta\hat{g}^t + e^t\|^2 \le 2\eta^2\|\hat{g}^t\|^2 + 2\|e^t\|^2.$$

Substituting these back yields:

$$\mathbb{E}[f^{t+1}] \le f^t - \eta\langle\nabla f, h^t\rangle + \frac{\eta}{8}\|\nabla f\|^2 + L\eta^2\mathbb{E}\|\hat{g}^t\|^2 + \Big(\frac{2}{\eta} + L\Big)\mathbb{E}\|e^t\|^2.$$

Next, we apply the polarization identity $-\eta\langle\nabla f, h^t\rangle = -\frac{\eta}{2}\|\nabla f\|^2 - \frac{\eta}{2}\|h^t\|^2 + \frac{\eta}{2}\|\nabla f - h^t\|^2$ and bound the stochastic variance $\mathbb{E}\|\hat{g}^t\|^2 \le \|h^t\|^2 + \frac{\sigma^2}{N}$:

$$\mathbb{E}[f^{t+1}] \le f^t - \frac{3\eta}{8}\|\nabla f\|^2 + \frac{\eta}{2}\|\nabla f - h^t\|^2 + \Big(\frac{2}{\eta} + L\Big)\rho^2 + \frac{L\eta^2\sigma^2}{N} + \eta\Big(L\eta - \frac{1}{2}\Big)\|h^t\|^2.$$

For $\eta \le \frac{1}{2L}$, the factor $(L\eta - 1/2) \le 0$, allowing us to drop the $\|h^t\|^2$ term. Finally, we bound the gradient bias using Jensen's inequality and L-smoothness: $\|\nabla f - h^t\|^2 \le \frac{L^2}{N}\|U^t\|_F^2$. Rearranging to isolate the gradient gives:

$$\frac{3\eta}{8}\mathbb{E}\|\nabla f(\bar{\theta}^t)\|^2 \le f^t - \mathbb{E}[f^{t+1}] + \frac{\eta L^2}{2N}\mathbb{E}\|U^t\|_F^2 + \frac{L\eta^2\sigma^2}{N} + \Big(\frac{2}{\eta} + L\Big)\rho^2. \quad (14)$$

c) Step 3: Consensus Contraction: The consensus error $U^{t+1} = (I - J)\Theta^{t+1}$ satisfies:

$$U^{t+1} = (I - J)\widetilde{W}^t(U^t - \eta V^t) + (I - J)E^t.$$

We apply Young's inequality $\|A+B\|_F^2 \le (1+\alpha_1)\|A\|_F^2 + (1+1/\alpha_1)\|B\|_F^2$ with $\alpha_1 = p/2$. Since $\widetilde{W}^t$ contracts by $(1-p)$, we have $\|A\|_F^2 \le (1-p)\|U^t - \eta V^t\|_F^2$. Observe that $(1+p/2)(1-p) = 1 - p/2 - p^2/2 \le 1 - p/2$, and $1 + 1/\alpha_1 = 1 + 2/p \le 3/p$. Thus:

$$\mathbb{E}\|U^{t+1}\|_F^2 \le (1 - \tfrac{p}{2})\mathbb{E}\|U^t - \eta V^t\|_F^2 + \frac{3}{p}\mathbb{E}\|E^t\|_F^2.$$

We apply Young's inequality again to split the inner term: $\|U^t - \eta V^t\|_F^2 \le (1+\alpha_2)\|U^t\|_F^2 + (1+1/\alpha_2)\eta^2\|V^t\|_F^2$, setting $\alpha_2 = p/4$. Observe that $(1 - p/2)(1 + p/4) = 1 - p/4 - p^2/8 \le 1 - p/4$, and $(1 - p/2)(1 + 4/p) \le 4/p$. Using (13) for $\mathbb{E}\|E^t\|_F^2 \le N\rho^2$, we obtain:

$$\mathbb{E}\|U^{t+1}\|_F^2 \le (1 - \tfrac{p}{4})\mathbb{E}\|U^t\|_F^2 + \frac{4\eta^2}{p}\mathbb{E}\|V^t\|_F^2 + \frac{3N}{p}\rho^2. \quad (15)$$

d) Step 4: Combine: Summing (15) over $t = 0$ to $T - 1$ bounds the cumulative consensus error:

$$\sum_{t=0}^{T-1}\mathbb{E}\|U^t\|_F^2 \le \frac{16\eta^2}{p^2}\sum_{t=0}^{T-1}\mathbb{E}\|V^t\|_F^2 + \frac{12NT}{p^2}\rho^2.$$

Recall from Step 2 of Theorem 1 that $\mathbb{E}\|V^t\|_F^2 \le 2L^2\mathbb{E}\|U^t\|_F^2 + 4N\tau_{\text{scr}}^2 + N\sigma^2 + 4N\mathbb{E}\|\nabla f\|^2$. Substituting this into the sum:

$$\sum_{t=0}^{T-1}\mathbb{E}\|U^t\|_F^2 \le \frac{32\eta^2L^2}{p^2}\sum_{t=0}^{T-1}\mathbb{E}\|U^t\|_F^2 + \frac{16\eta^2N}{p^2}\sum_{t=0}^{T-1}\big(4\tau_{\text{scr}}^2 + \sigma^2 + 4\|\nabla f\|^2\big) + \frac{12NT}{p^2}\rho^2.$$

For $\eta \le \frac{p}{8L}$, we have $\frac{32\eta^2L^2}{p^2} \le \frac{1}{2}$. Moving the U^t sum to the left and multiplying by 2:

$$\sum_{t=0}^{T-1}\mathbb{E}\|U^t\|_F^2 \le \frac{32\eta^2N}{p^2}\sum_{t=0}^{T-1}\big(4\tau_{\text{scr}}^2 + \sigma^2 + 4\|\nabla f(\bar{\theta}^t)\|^2\big) + \frac{24NT}{p^2}\rho^2.$$

We now substitute this consensus sum into the telescoped Descent Lemma (14):

$$\frac{3\eta}{8}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f\|^2 \le (f^0 - f^\star) + \frac{L\eta^2\sigma^2T}{N} + \Big(\frac{2}{\eta} + L\Big)\rho^2T + \frac{64\eta^3L^2T\tau_{\text{scr}}^2}{p^2} + \frac{16\eta^3L^2T\sigma^2}{p^2} + \frac{12\eta L^2T}{p^2}\rho^2 + \frac{64\eta^3L^2}{p^2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f\|^2.$$

To absorb the gradient term into the left-hand side, we require its coefficient to be $\le \frac{\eta}{8}$, leaving $\frac{\eta}{4}$ on the left. This requires $\frac{64\eta^3L^2}{p^2} \le \frac{\eta}{8}$, which simplifies to $\eta^2 \le \frac{p^2}{512L^2}$, yielding the stepsize condition $\eta \le \frac{p}{24L}$.

Assuming this stepsize, subtracting the gradient term leaves $\frac{\eta}{4}\sum\mathbb{E}\|\nabla f\|^2$ on the left. Dividing the entire inequality by $\frac{\eta T}{4}$ yields the final rigorous bound:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\theta}^t)\|^2 \le \frac{4(f^0 - f^\star)}{\eta T} + \frac{4L\eta\sigma^2}{N} + \frac{256L^2\eta^2(\tau_{\text{scr}}^2 + \sigma^2/4)}{p^2} + \frac{48L^2}{p^2}\rho^2 + \Big(\frac{8}{\eta^2} + \frac{4L}{\eta}\Big)\rho^2.$$

□

VIII. Distributed-Backend Verification

To confirm that the simulation backend of Section V (main paper) faithfully reproduces real distributed execution, we ran DMTT and FedAvg (dynamic) on the ZeroMQ distributed backend at 10 nodes across three Byzantine fractions (10%, 30%, 50%) and two seeds each (20 rounds, $\Delta = 10$ s per round, IPC transport). Figure 1 overlays the per-round honest-node accuracy from both backends.

Across all six (condition, fraction) pairs the two backends produce qualitatively identical trajectories. For DMTT, the mean final honest-node accuracy (last 5 of 20 rounds) agrees within 0.07 between the two backends across all fractions; the largest gap occurs at 30% Byzantine (0.845 distributed vs. 0.775 simulation), attributable to the short 20-round window over which DMTT is still mid-convergence. For FedAvg (dynamic), both backends agree that the unprotected baseline collapses to near-chance (0.13–0.18) regardless of Byzantine fraction. This qualitative agreement confirms that coordinator-free wall-clock synchronisation and peer-to-peer ZeroMQ transport introduce no systematic deviation from the in-process simulation, justifying the use of the simulation backend for the main 100-node sweep in Section VII.

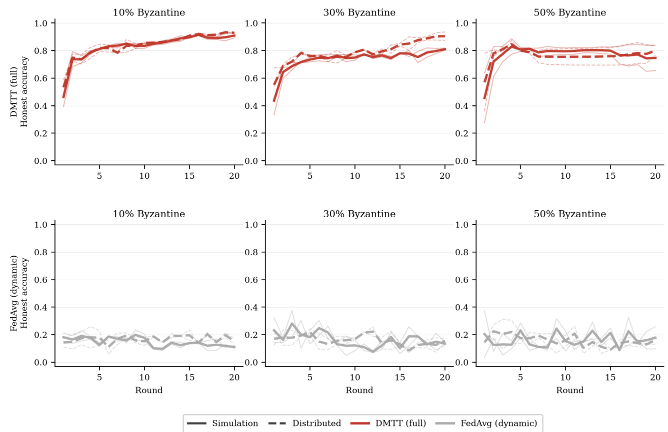

Fig. 1. Distributed-backend verification at 10 nodes. Solid lines: simulation backend; dashed lines: ZeroMQ distributed backend (real OS processes, wall-clock synchronisation). Thin lines show individual seeds; bold lines show the mean over 2 seeds. Rows: DMTT (top) and FedAvg dynamic (bottom). Columns: 10%, 30%, 50% Byzantine fraction.

REFERENCES

[1] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *International conference on machine learning*. PMLR, 2020, pp. 5381–5393.