SketchGuard: Scaling Byzantine-Robust Decentralized Federated Learning via Sketch-Based Screening

Supplemental Material

Murtaza Rangwala, Farag Azzedin, Richard O. Sinnott, and Rajkumar Buyya

This document provides supplemental material for the main paper. It contains: (S1) full proofs of Theorems 1 and 2 from the main paper; (S2) detailed model architecture specifications; (S3) network topology specifications; (S4) complete hyperparameter settings; and (S5) backdoor trigger specifications and scalability experiment configurations.

All notation is consistent with the main paper. We use $\mathcal{H}$ to denote the set of honest clients, $\mathcal{S}_i^t$ for the set of accepted neighbors at round t, and $\gamma_{\text{eff}} = \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}$ for the effective filtering threshold under sketch approximation parameter ϵ.

I. Proof of Theorem 1: Strongly Convex Convergence

We prove convergence for SketchGuard under μ-strongly convex, L-smooth objectives. The proof follows the analysis framework of BALANCE [1] and modifies it to handle Count Sketch compression via the effective threshold γ_{eff}.

Remark I.1 (Role of graph topology in the proof). The convergence bound is stated and proved for each benign client i individually, tracking $\mathbb{E}[F(\mathbf{w}_i^T) - F(\mathbf{w}^*)]$ for client i's own model. This is the same proof structure as BALANCE [1]. The analysis does *not* require a graph mixing or spectral gap argument, because it does not track a network-average or consensus model. The communication graph enters the proof only through Assumption 5 (connectivity of $G_{\mathcal{H}}$), which ensures no honest client is permanently isolated from honest information. The per-round neighbor-deviation bound (8) absorbs the local aggregation error directly; error does not accumulate across rounds via graph propagation because each round's aggregation is self-contained.

SketchGuard's sketch modification affects only the *filtering criterion* used to construct $\mathcal{S}_i^t$. It does not change the aggregation weights, the graph structure, or any other part of the protocol. The graph topology argument (Assumption 5) is therefore inherited verbatim from BALANCE [1], and the sketch compression does not perturb it in any way.

A. Probability Event and Notation

We condition throughout on the event $\mathcal{E}$ that all pairwise sketch distance comparisons simultaneously satisfy the distance preservation guarantee of Lemma 1 (main paper). By a union bound over all T rounds, n nodes, and maximum degree $\Delta = \max_i |\mathcal{N}_i|$, the per-comparison failure probability ζ must satisfy $T \cdot n \cdot \Delta \cdot \zeta \leq \zeta_{\text{sys}}$. Setting $\zeta = \zeta_{\text{sys}}/(T \cdot n \cdot \Delta)$ yields $\Pr[\mathcal{E}] \geq 1 - \zeta_{\text{sys}}$, and corresponds to the sketch size $k = O(\epsilon^{-2} \log(Tn\Delta/\zeta_{\text{sys}}))$ stated in the theorem.

We write $\mathbf{w}_i^t$ for client i's model at the start of round t, $\mathbf{w}_i^{t+1/2} = \mathbf{w}_i^t - \eta \mathbf{g}(\mathbf{w}_i^t)$ for the half-updated model after local training, and drop the round superscript on $\mathcal{S}_i^t$ writing it as $\mathcal{S}_i$ for brevity within this proof.

B. Model Update Decomposition

The full aggregated update from round t to $t+1$ is:

$$\mathbf{w}_i^{t+1} - \mathbf{w}_i^t = -\eta \mathbf{g}(\mathbf{w}_i^t) + \frac{1-\alpha}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \big(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2}\big). \tag{1}$$

The first term is the local gradient step; the second is the correction from neighbor aggregation.

C. Applying L-Smoothness

By L-smoothness of F (Assumption 2, main paper):

$$F(\mathbf{w}_i^{t+1}) \leq F(\mathbf{w}_i^t) + \langle \nabla F(\mathbf{w}_i^t),\, \mathbf{w}_i^{t+1} - \mathbf{w}_i^t \rangle + \frac{L}{2} \|\mathbf{w}_i^{t+1} - \mathbf{w}_i^t\|^2. \tag{2}$$

Substituting the update decomposition:

$$F(\mathbf{w}_i^{t+1}) \le F(\mathbf{w}_i^t) - \eta\langle \nabla F(\mathbf{w}_i^t), \mathbf{g}(\mathbf{w}_i^t)\rangle + \Big\langle \nabla F(\mathbf{w}_i^t),\, \frac{1-\alpha}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\rangle + \frac{L}{2}\Big\| -\eta\mathbf{g}(\mathbf{w}_i^t) + \frac{1-\alpha}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\|^2. \tag{3}$$

D. Bounding the Quadratic Term

Using $\|\mathbf{a}+\mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$:

$$\Big\| -\eta\mathbf{g}(\mathbf{w}_i^t) + \frac{1-\alpha}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\|^2 \le 2\eta^2\|\mathbf{g}(\mathbf{w}_i^t)\|^2 + 2\Big\|\frac{1-\alpha}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\|^2. \tag{4}$$

E. Impact of Sketch-Based Filtering

Conditioned on event $\mathcal{E}$, for any $j \in \mathcal{S}_i$ (accepted by the sketch filter):

$$\|\mathrm{CS}(\mathbf{w}_i^{t+1/2}) - \mathrm{CS}(\mathbf{w}_j^{t+1/2})\| \le \gamma \exp\Big(-\frac{\kappa t}{T}\Big)\|\mathrm{CS}(\mathbf{w}_i^{t+1/2})\|. \tag{5}$$

Applying the right-hand side of the distance preservation lemma to the left-hand side and the left-hand side to the right-hand side:

$$(1-\epsilon)^{1/2}\|\mathbf{w}_i^{t+1/2} - \mathbf{w}_j^{t+1/2}\| \le \gamma \exp\Big(-\frac{\kappa t}{T}\Big)(1+\epsilon)^{1/2}\|\mathbf{w}_i^{t+1/2}\|. \tag{6}$$

Rearranging:

$$\|\mathbf{w}_i^{t+1/2} - \mathbf{w}_j^{t+1/2}\| \le \gamma_{\text{eff}} \exp\Big(-\frac{\kappa t}{T}\Big)\|\mathbf{w}_i^{t+1/2}\|, \tag{7}$$

where $\gamma_{\text{eff}} = \gamma\sqrt{(1+\epsilon)/(1-\epsilon)}$. Averaging over $\mathcal{S}_i$ and applying the bounded-parameters assumption $\|\mathbf{w}_i^{t+1/2}\| \le \psi$:

$$\Big\|\frac{1}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\| \le \gamma_{\text{eff}}\,\psi. \tag{8}$$

F. Bounding the Inner Product Terms

For the gradient inner product, using unbiasedness of $\mathbf{g}$ (Assumption 3, main paper):

$$\mathbb{E}[-\eta\langle \nabla F(\mathbf{w}_i^t), \mathbf{g}(\mathbf{w}_i^t)\rangle] = -\eta\|\nabla F(\mathbf{w}_i^t)\|^2. \tag{9}$$

For the neighbor inner product, applying Cauchy–Schwarz and the gradient bound $\|\nabla F(\mathbf{w}_i^t)\| \le \rho$ (Assumption 4, main paper):

$$\Big\langle \nabla F(\mathbf{w}_i^t),\, \frac{1-\alpha}{|\mathcal{S}_i|}\sum_{j\in\mathcal{S}_i}(\mathbf{w}_j^{t+1/2} - \mathbf{w}_i^{t+1/2})\Big\rangle \le \|\nabla F(\mathbf{w}_i^t)\| \cdot (1-\alpha)\gamma_{\text{eff}}\psi \le \rho(1-\alpha)\gamma_{\text{eff}}\psi. \tag{10}$$

G. Combining and Taking Expectations

Using $\mathbf{g} = (\mathbf{g} - \nabla F) + \nabla F$ and $\|\mathbf{a}+\mathbf{b}\|^2 \le 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$:

$$\mathbb{E}[\|\mathbf{g}(\mathbf{w}_i^t)\|^2] \le 2\|\nabla F(\mathbf{w}_i^t)\|^2 + 2\delta^2. \tag{11}$$

Substituting into the smoothness bound and taking full expectation:

$$\mathbb{E}[F(\mathbf{w}_i^{t+1})] \le \mathbb{E}[F(\mathbf{w}_i^t)] + (-\eta + 2L\eta^2)\|\nabla F(\mathbf{w}_i^t)\|^2 + 2L\eta^2\delta^2 + 2L\eta^2\big(\gamma_{\text{eff}}\psi(1-\alpha)\big)^2 + \rho(1-\alpha)\gamma_{\text{eff}}\psi. \tag{12}$$

The quadratic neighbor term $2L\eta^2(\gamma_{\text{eff}}\psi(1-\alpha))^2$ is of order η^2 and dominated by the linear term $\rho(1-\alpha)\gamma_{\text{eff}}\psi$ for small η; we absorb it into a single constant. Applying $\eta \le 1/(4L)$, which gives $-\eta + 2L\eta^2 \le -\eta/2$:

$$\mathbb{E}[F(\mathbf{w}_i^{t+1})] \le \mathbb{E}[F(\mathbf{w}_i^t)] - \frac{\eta}{2}\|\nabla F(\mathbf{w}_i^t)\|^2 + 2L\eta^2\delta^2 + 2\gamma_{\text{eff}}\psi\rho(1-\alpha). \tag{13}$$

H. Applying Strong Convexity via the PL Inequality

Strong convexity implies the Polyak–Łojasiewicz (PL) inequality:

$$\|\nabla F(\mathbf{w}_i^t)\|^2 \geq 2\mu\big(F(\mathbf{w}_i^t) - F(\mathbf{w}^*)\big). \tag{14}$$

Substituting into (13) and rearranging:

$$\mathbb{E}[F(\mathbf{w}_i^{t+1}) - F(\mathbf{w}^*)] \leq (1-\mu\eta)\,\mathbb{E}[F(\mathbf{w}_i^t) - F(\mathbf{w}^*)] + 2L\eta^2\delta^2 + 2\gamma_{\text{eff}}\psi\rho(1-\alpha). \tag{15}$$

I. Telescoping

Applying the above recursion from $t=0$ to $T-1$ and using $\eta \leq 1/\mu$ to ensure $1-\mu\eta \in [0,1)$:

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_i^T) - F(\mathbf{w}^*)] &\leq (1-\mu\eta)^T\big[F(\mathbf{w}_i^0) - F(\mathbf{w}^*)\big] \\ &\quad + \frac{2L\eta^2\delta^2 + 2\gamma_{\text{eff}}\psi\rho(1-\alpha)}{1-(1-\mu\eta)} \\ &\leq (1-\mu\eta)^T\big[F(\mathbf{w}_i^0) - F(\mathbf{w}^*)\big] + \frac{2L\eta\delta^2}{\mu} + \frac{2\gamma_{\text{eff}}\rho\psi(1-\alpha)}{\mu\eta}.\end{aligned} \tag{16}$$

This completes the proof of Theorem 1. The compression error enters only through γ_{eff}; setting $\epsilon \to 0$ recovers the BALANCE full-precision bound [1]. □

II. Proof of Theorem 2: Non-Convex Convergence

We establish the non-convex convergence guarantee, adapting the one-step bound derived above.

A. One-Step Bound

The one-step bound (13) holds without any convexity assumption—it follows from L-smoothness alone. Rearranging:

$$\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_i^t)\|^2] \leq \mathbb{E}[F(\mathbf{w}_i^t) - F(\mathbf{w}_i^{t+1})] + 2L\eta^2\delta^2 + 2\gamma_{\text{eff}}\psi\rho(1-\alpha). \tag{17}$$

B. Telescoping and Averaging

Summing from $t=0$ to $T-1$:

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\mathbf{w}_i^t)\|^2] \leq \mathbb{E}[F(\mathbf{w}_i^0) - F(\mathbf{w}_i^T)] + T\cdot(2L\eta^2\delta^2 + 2\gamma_{\text{eff}}\psi\rho(1-\alpha)). \tag{18}$$

Using $F(\mathbf{w}_i^T) \geq F^* = \inf_{\mathbf{w}} F(\mathbf{w})$ and dividing by $\eta T/2$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F(\mathbf{w}_i^t)\|^2] \leq \frac{2[F(\mathbf{w}_i^0) - F^*]}{\eta T} + 4L\eta\delta^2 + \frac{4\gamma_{\text{eff}}\psi\rho(1-\alpha)}{\eta}. \tag{19}$$

This is the statement of Theorem 2. As in the strongly convex case, setting $\epsilon \to 0$ recovers the BALANCE non-convex bound [1]. □

III. Model Architecture Specifications

A. FEMNIST

The FEMNIST model follows the original LEAF specification: two convolutional layers with max-pooling, followed by two fully connected layers. The architecture's parameter count $d = 6{,}603{,}646$ seeds the Count Sketch hash functions. Table I gives the full specification.

TABLE I
FEMNIST model architecture ($d = 6{,}603{,}646$).

Layer	Type	Output Shape	Parameters
Input	–	$1 \times 28 \times 28$	–
Conv1	Conv2d	$32 \times 28 \times 28$	832
Pool1	MaxPool2d	$32 \times 14 \times 14$	–
Conv2	Conv2d	$64 \times 14 \times 14$	51,264
Pool2	MaxPool2d	$64 \times 7 \times 7$	–
Flatten	–	3,136	–
FC1	Linear	2,048	6,424,576
FC2	Linear	62	126,974
Total			**6,603,646**

B. CelebA

The CelebA model uses a LeNet-style architecture adapted for 84×84 RGB input and binary classification. Table II gives the full specification.

TABLE II
CELEBA MODEL ARCHITECTURE ($d = 2{,}219{,}692$).

Layer	Type	Output Shape	Parameters
Input	–	$3 \times 84 \times 84$	–
Conv1	Conv2d	$30 \times 84 \times 84$	840
Pool1	MaxPool2d	$30 \times 42 \times 42$	–
Conv2	Conv2d	$50 \times 42 \times 42$	13,550
Pool2	MaxPool2d	$50 \times 21 \times 21$	–
Flatten	–	22,050	–
FC1	Linear	100	2,205,100
FC2	Linear	2	202
Total			**2,219,692**

C. Sent140

The Sent140 model uses a stacked LSTM architecture following the LEAF reference implementation. Table III gives the full specification.

TABLE III
SENT140 MODEL ARCHITECTURE ($d \approx 1.2$M).

Layer	Type	Configuration	Output
Input	–	–	seq_len
Embedding	Embedding	vocab $= 10{,}000$, dim $= 100$	seq $\times$ 100
LSTM	2-layer	hidden $= 100$	100
Dropout	Dropout	$p = 0.1$	100
FC1	Linear	$100 \to 128$	128
FC2	Linear	$128 \to 2$	2
Total			**$\approx$1.2M**

IV. NETWORK TOPOLOGY SPECIFICATIONS

Table IV specifies parameters and expected node degrees for each topology used in the robustness experiments (20-node networks). The Erdős-Rényi topologies resample edges each round to model realistic ad hoc connectivity with intermittent connections.

TABLE IV
NETWORK TOPOLOGY CONFIGURATIONS (20-NODE NETWORKS).

Topology	Parameter	Expected Degree
Ring	–	2
ER (sparse)	$p = 0.20$	3.8
ER (medium)	$p = 0.45$	8.6
ER (dense)	$p = 0.60$	11.4
Fully Connected	–	19

V. COMPLETE HYPERPARAMETER SETTINGS

Table V provides the complete hyperparameter configuration used across all experiments.

TABLE V
COMPLETE HYPERPARAMETER CONFIGURATION.

Parameter	Value
Training Configuration	
Number of clients	20
Global rounds	50
Local epochs per round	3
Batch size	64
Learning rate	0.01 (FEMNIST, Sent140); 0.001 (CelebA)
Max samples per client	10,000 (FEMNIST); 4,500 (CelebA, Sent140)
Random seeds	987654321; 39573295; 32599368
SketchGuard Parameters	
Sketch size (k)	1,000 (FEMNIST); 350 (CelebA); 180 (Sent140)
Threshold (γ)	2.0
Decay (κ)	1.0
Mixing (α)	0.5
Hash seed	42
BALANCE Parameters	
Threshold (γ)	2.0
Decay (κ)	1.0
Mixing (α)	0.5
UBAR Parameters	
Robustness (ρ)	$1.0 -$ attack fraction
KRUM Parameters	
Compromised (f)	$\lfloor$attack fraction $\times\ \lvert\mathcal{N}_i\rvert\rfloor$

VI. BACKDOOR ATTACK AND SCALABILITY EXPERIMENT SPECIFICATIONS

A. Backdoor Trigger Specifications

Trigger patterns are chosen to be visually or lexically distinctive from natural inputs and are absent from the clean test set, so that ASR meaningfully measures the backdoor's presence.

FEMNIST: A white pixel square of size 4×4 pixels placed in the bottom-right corner of each input image. Target label: class 0 (digit "0").

CelebA: A white pixel square of size 8×8 pixels placed in the bottom-right corner of each input image. Target label: class 0 (non-smiling).

Sent140: Three rare vocabulary tokens (indices 9997, 9998, 9999) appended to the end of each input sequence. These indices correspond to out-of-distribution tokens that rarely appear in natural English text. Target label: class 1 (positive sentiment).

In all cases, Byzantine clients apply a scaling factor γ_{bd} to their model updates before submission, following the model-replacement methodology of [2], with γ_{bd} set to overcome honest averaging in each experimental setting.

B. Network Size Scaling Configurations

Table VI describes the k-regular graph configurations used in the network size scaling experiments. All configurations maintain 50% Byzantine clients. Experiments use 3 training rounds with 1 local epoch per round to isolate computational overhead from training dynamics. All wall-clock measurements are recorded on CPU hardware.

TABLE VI
NETWORK SIZE SCALING EXPERIMENT CONFIGURATIONS.

Node Degree	Network Size	Byzantine %
16	20	50%
32	35	50%
96	100	50%
154	155	50%
299	300	50%

C. Model Size Scaling Variants

Table VII describes the five FEMNIST architecture variants used in the model size scaling experiments. All variants preserve the two-convolutional-layer, two-fully-connected-layer structure; only filter counts and layer widths are scaled. All model scaling experiments use a fixed k-regular graph with node degree 154 (155 total nodes) to isolate model size effects.

TABLE VII
FEMNIST MODEL SCALING VARIANTS.

Variant	Modification	Parameters
Tiny	Reduced filters and hidden units	220,318
Small	Standard configuration	848,382
Baseline	Baseline FEMNIST architecture	6,603,710
Large	Increased filters and hidden units	26,154,814
XLarge	Further increased dimensions	60,271,678

REFERENCES

[1] M. Fang, Z. Zhang, Hairi, P. Khanduri, J. Liu, S. Lu, Y. Liu, and N. Gong, "Byzantine-robust decentralized federated learning," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '24. Association for Computing Machinery, 2024, p. 2874–2888.

[2] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 108. PMLR, 26–28 Aug 2020, pp. 2938–2948.