ON JOHNSON-LINDENSTRAUSS TRANSFORMS FOR BYZANTINE-ROBUST OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The growing scale of data and models in machine learning calls for efficient optimization methods. Distributed optimization addresses this challenge, with compression mechanisms helping to reduce communication costs. A classical tool for dimensionality reduction is the Johnson-Lindenstrauss (JL) transform, which is simple to implement and offers strong geometric guarantees. In particular, JL transforms approximately preserve inner products and pairwise distances between vectors with high probability. In this paper, we exploit this property in Byzantine-robust optimization, where worker messages are weighted according to their alignment with a trusted reference gradient. We further support our theoretical findings with experimental results.

1 INTRODUCTION

The growing scale of data and computational demands of machine learning (Kairouz et al., 2021) have motivated distributed optimization methods (Tang et al., 2020) for the global objective

$$\min_{x\in\mathbb{R}^n}\left\{f(x)=\frac{1}{G}\sum_{i\in\mathcal{G}}f_i(x)\right\}, \tag{1}$$

where f_i denotes the local loss at worker i, $\mathcal{G}\subseteq\{1,\dots,M\}$ is the fixed set of honest workers, and $G=|\mathcal{G}|\geq 1$. The objective is the average loss over the honest workers, whose identities need not be known to the server; when all workers are honest, it reduces to the usual average over all M workers. Distributed methods make large-scale training tractable when a single machine is insufficient, but introduce communication bottlenecks, synchronization overhead, and bandwidth limitations that can substantially hinder performance (Li et al., 2019).

To mitigate the communication cost, a broad range of compression techniques has been proposed (McMahan et al., 2017; Alistarh et al., 2017), such as quantization (Yu et al., 2019), which lowers numerical precision, and sparsification (Wangni et al., 2018), which communicates only the most informative coordinates. Workers transmit either compressed gradients (Alistarh et al., 2017) or auxiliary sequences that compensate for compression errors across iterations (Karimireddy et al., 2019; Richtárik et al., 2021; Mishchenko et al., 2024). In both cases, the gradient estimator is obtained by averaging compressed vectors.

However, some distributed setups rely not on the mean itself, but on geometric relations between gradients. A prominent example is Byzantine-robust optimization (Karimireddy et al., 2021), in which Byzantine workers attempt to derail training by sending arbitrary adversarial messages. One remedy is to assign reliability weights to workers (Cao et al., 2020), computed from inner products of local gradients with a trusted reference gradient (Yan et al., 2023; Molodtsov et al., 2025), so that workers not aligned with the reference direction are suppressed:

$$\omega_i^t=\frac{\max\left\{\langle\nabla f_i(x^t),\nabla f_0(x^t)\rangle,0\right\}}{\sum_{j=1}^{M}\max\left\{\langle\nabla f_j(x^t),\nabla f_0(x^t)\rangle,0\right\}},\quad x^{t+1}=x^t-\gamma\sum_{i=1}^{M}\omega_i^t\nabla f_i(x^t), \tag{2}$$

where f_0 is a loss function defined on a trusted device. In this illustrative update, the iterate is left unchanged if the denominator is zero, and Byzantine workers may replace their gradients by arbitrary messages.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Such methods depend on preserving ℓ_2-geometry rather than exact coordinate values, which motivates compression operators that approximately preserve Euclidean structure. Johnson–Lindenstrauss (JL) transforms (Johnson et al., 1984) map high-dimensional vectors into a lower-dimensional space while approximately preserving pairwise distances and inner products. We focus on stochastic JL transforms, as they are simple to implement in practice. They have been successfully applied in nearest-neighbor search (Kushilevitz et al., 1998; Andoni et al., 2015), random partitioning (Kleinberg, 1997), linear regression (Heinze et al., 2016; Thanei et al., 2017), clustering (Tang et al., 2005; Boutsidis et al., 2010; Dasgupta, 1999; Urruty et al., 2007; Heckel et al., 2017; Guo et al., 2020), differential privacy (Upadhyay, 2014; Sheffet, 2017), and numerical linear algebra (Woodruff et al., 2014; Musco & Musco, 2020; Sarlos, 2006).

In distributed optimization and federated learning, however, JL transforms have mostly been treated as unbiased sketching operators (Song et al., 2021; Shrivastava et al., 2024; Dai et al., 2019), leading to guarantees in expectation, or derandomized (Acharya et al., 2019). These works do not capitalize on the preservation of distances and inner products, and thus do not clearly distinguish JL transforms from standard sparsification or sketching schemes.

Our contributions. In this paper, we

- Propose a JL-based communication mechanism for Byzantine-robust distributed optimization methods that rely on gradient alignment with a trusted reference gradient;
- Use high-probability control of compressed honest gradients to derive convergence guarantees without convexity, under average data similarity and an explicit accuracy condition on the trusted reference gradient;
- Validate our theoretical findings experimentally and demonstrate the practical effectiveness of the proposed approach.

2 Proposed Algorithm

2.1 Notation

We equip $\mathbb{R}^n$ with the standard Euclidean norm, $\|x\| \stackrel{\text{def}}{=} \langle x, x\rangle^{1/2}, x \in \mathbb{R}^n$. Throughout the paper, $f : \mathbb{R}^n \to \mathbb{R}$ denotes a differentiable objective function.

Assumption 2.1 (Smoothness)**.** Differentiable function f_0 has L-Lipschitz gradient for some $L > 0$, i.e.

$$\|\nabla f_0(x) - \nabla f_0(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n. \tag{3}$$

Moreover, $f_0^{\inf} \stackrel{\text{def}}{=} \inf_{x\in\mathbb{R}^n} f_0(x) > -\infty$.

2.2 JL Transforms

Definition 2.2 (JL Transform (Johnson et al., 1984))**.** Given $\varepsilon, \delta \in (0, 1)$, a stochastic mapping $h : \mathbb{R}^n \to \mathbb{R}^k$ is called a JL transform if, for all $u, v \in \mathbb{R}^n$,

$$\mathbb{P}\left[(1-\varepsilon)\|u - v\|^2 \leq \|h(u) - h(v)\|^2 \leq (1+\varepsilon)\|u - v\|^2\right] \geq 1 - \delta. \tag{4}$$

In words, a JL transform preserves all pairwise distances up to a factor $1 \pm \varepsilon$ with probability at least $1 - \delta$, while reducing the dimension from n to k. If h is linear, then $h(0) = 0$, and taking $v = 0$ shows that norms are preserved as well: $(1-\varepsilon)\|u\|^2 \leq \|h(u)\|^2 \leq (1+\varepsilon)\|u\|^2$. Since $\langle u, v\rangle = \frac{1}{4}\left(\|u + v\|^2 - \|u - v\|^2\right)$, preserving the norms of $u + v$ and $u - v$ also preserves inner products.

Lemma 2.3. *If h is a linear JL Transform, then $\mathbb{P}\left[\,|\,\langle h(u), h(v)\rangle - \langle u, v\rangle\,|\leq \varepsilon\|u\|\,\|v\|\,\right] \geq 1 - 2\delta$ for all $u, v \in \mathbb{R}^n$.*

We consider two standard linear constructions.

Definition 2.4 (Gaussian JL matrix)**.** $S \in \mathbb{R}^{k\times n}$ with i.i.d. entries $[S]_{ij} \sim \mathcal{N}(0, \frac{1}{k})$.

Definition 2.5 (Rademacher JL matrix)**.** $S \in \mathbb{R}^{k\times n}$ with i.i.d. entries $[S]_{ij} = \pm\frac{1}{\sqrt{k}}$, each with probability $\frac{1}{2}$.

Both are JL transforms with distortion $\varepsilon \in (0,1)$ and failure probability $\delta \in (0,1)$ whenever $k \geq C\varepsilon^{-2}\log(2/\delta)$, where $C > 0$ is a universal constant (Dasgupta & Gupta, 2003; Achlioptas, 2003). To return to $\mathbb{R}^n$ for the optimization update, we use the lifted operator $\mathrm{JL}_S(x) \overset{\text{def}}{=} S^\top S x \in \mathbb{R}^n$.

Lemma 2.6 (Unbiasedness). *For S as in Definitions 2.4 and 2.5, $\mathbb{E}\,\mathrm{JL}_S(x) = x$ for all $x \in \mathbb{R}^n$.*

Lemma 2.7 (Second moment). *For all $x \in \mathbb{R}^n$, $\mathbb{E}\|\mathrm{JL}_S(x)\|^2 = \frac{n+k+1}{k}\|x\|^2$ for Gaussian and $\mathbb{E}\|\mathrm{JL}_S(x)\|^2 = \frac{n+k-1}{k}\|x\|^2$ for Rademacher S.*

Thus JL_S resembles unbiased sparsification such as `Randk`, $\mathcal{C}(x) = \frac{n}{k}\sum_{i\in I} x_i e_i$ with $I \subset [n]$ a uniformly random subset of size k, which is also unbiased with variance of order $\frac{n}{k}$ (Beznosikov et al., 2023). The key distinction is that the transmitted message of `Randk`, $h_{\mathcal{C}}(x) \overset{\text{def}}{=} \sqrt{n/k}\,(x_i)_{i\in I} \in \mathbb{R}^k$ (so that $\mathcal{C} = h_{\mathcal{C}}^\top h_{\mathcal{C}}$, just as $\mathrm{JL}_S = S^\top S$), does not preserve norms with high probability. By linearity, the JL property for $h_{\mathcal{C}}$ reduces to $\mathbb{P}\big[\big|\|h_{\mathcal{C}}(x)\|^2/\|x\|^2 - 1\big| \leq \varepsilon\big] \geq 1-\delta$ for all $x \in \mathbb{R}^n \setminus \{0\}$. For $x = e_1$ the ratio is 0 with probability $1 - \frac{k}{n}$, forcing $\delta \geq 1 - \frac{k}{n}$.

Lemma 2.8 (Operator-norm bound (Vershynin, 2026)). *For S as in Definitions 2.4 or 2.5, there exists a universal constant $c > 0$, independent of n and k, such that, with probability at least $1-\delta$,*

$$\|S\| \leq 1 + c\sqrt{\frac{n}{k}} + c\sqrt{\frac{\log(2/\delta)}{k}}. \tag{5}$$

Consequently, on the same event, $\|\mathrm{JL}_S(x)\|^2 \leq \left(1 + c\sqrt{\frac{n}{k}} + c\sqrt{\frac{\log(2/\delta)}{k}}\right)^4 \|x\|^2$ for every $x \in \mathbb{R}^n$.

The operator norm controls the lifted update, while the JL property controls the geometry of fixed vectors in the compressed space; the high-probability analysis uses both.

2.3 GRAD-BANT

Many Byzantine-resilient methods exploit gradient similarity to distinguish honest workers from adversarial ones. We introduce `Grad-BANT` (Algorithm 1), inspired by (Yan et al., 2023). As in equation 2, it relies on a trial function f_0 stored on the trusted server, whose gradient $\nabla f_0(x)$ is used to score the received messages; its accuracy with respect to the honest objective is specified in Assumption 2.10.

Algorithm 1 `Grad-BANT`

Parameters: $x^0 \in \mathbb{R}^n$, $\omega_i^{-1} = 1/M$, stepsize $\gamma > 0$, momentum $\beta \in (0,1]$, smoothness constant $L > 0$
for $t = 0, 1, \ldots, T-1$ **do**
  Generate a fresh S_t; server sends x^t and the sketch seed to workers
  for all $i = 1, \ldots, M$ in parallel **do**
    Honest worker: $g_i^t = \nabla f_i(x^t)$; send $y_i^t = S_t g_i^t$ to server
  end for
  $u_i^t = S_t^\top y_i^t$; $\theta_i^t = \langle S_t \nabla f_0(x^t), y_i^t\rangle - \frac{L\gamma}{2}\|u_i^t\|^2$
  if all $[\theta_i^t]_0 = 0$ **then**
    $\omega_i^t = (1-\beta)\omega_i^{t-1} + \frac{\beta}{M}$
  else
    $\omega_i^t = (1-\beta)\omega_i^{t-1} + \beta\dfrac{[\theta_i^t]_0}{\sum_{j=1}^M [\theta_j^t]_0}$
  end if
  $x^{t+1} = x^t - \gamma S_t^T \sum_{i=1}^M \mathbb{I}_{[\theta_i^t > 0]}\omega_i^t y_i^t$
end for

At each iteration, an honest worker sends $y_i^t = S_t g_i^t$ with $g_i^t = \nabla f_i(x^t)$, while a Byzantine worker may send any finite, measurable $y_i^t \in \mathbb{R}^k$, possibly chosen after observing the current sketch and the honest messages; $u_i^t = S_t^\top y_i^t$ denotes the lifted message. Each sketch is sampled independently of

all data and previous randomness after x^t is determined, so the honest gradients and the reference gradient are fixed conditional on the history before sampling S_t.

The trust score θ_i^t measures how well the lifted message aligns with the trusted reference gradient $\nabla f_0(x^t)$, while accounting for its norm: if $\langle \nabla f_0(x^t), u_i^t \rangle > \frac{L\gamma}{2}\|u_i^t\|^2$, the direction is sufficiently aligned with the trusted descent direction and receives a positive weight; otherwise, it is suppressed by the indicator in the update. The scores are clipped by $[x]_0 \overset{\text{def}}{=} \max\{x, 0\}$ and normalized. If all scores are non-positive, we blend the previous weights with uniform weights (line 9) and leave the iterate unchanged. The momentum parameter β smooths the evolution of the weights over time.

For honest workers, the server thus uses the compressed alignment terms $\langle S_t \nabla f_0(x^t), S_t g_i^t \rangle$, which the JL property keeps close to the exact ones with high probability, while the norm penalty is evaluated on the actual lifted message. The aggregated update is lifted back to $\mathbb{R}^n$ via $S_t^\top$ (line 13). Our analysis uses the aggregate honest alignment and the sum of honest squared norms; it does not require the compressed and uncompressed scores to select the same workers. We require the following similarity condition.

Assumption 2.9 (Data similarity)**.** The honest workers are said to satisfy (δ_1, δ_2)-data similarity if there exist $\delta_1, \delta_2 \geq 0$ such that

$$\frac{1}{G}\sum_{i\in\mathcal{G}} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \delta_1 + \delta_2 \|\nabla f(x)\|^2, \qquad \forall x \in \mathbb{R}^n. \tag{6}$$

This condition bounds heterogeneity on average over the honest workers, which is weaker than imposing it on every honest worker. Since f is their average, it also gives

$$\frac{1}{G}\sum_{i\in\mathcal{G}} \|\nabla f_i(x)\|^2 = \|\nabla f(x)\|^2 + \frac{1}{G}\sum_{i\in\mathcal{G}} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq (1+\delta_2)\|\nabla f(x)\|^2 + \delta_1. \tag{7}$$

The next assumption quantifies how accurately the trusted reference represents the objective; η_0 and η_1 capture the mismatch between the trial and the true data distributions.

Assumption 2.10 (Reference accuracy)**.** There exist $\eta_0 \geq 0$ and $\eta_1 \in [0, 1)$ such that

$$\|\nabla f_0(x) - \nabla f(x)\|^2 \leq \eta_0 + \eta_1 \|\nabla f(x)\|^2, \qquad \forall x \in \mathbb{R}^n. \tag{8}$$

Theorem 2.11. *Suppose that Assumptions 2.1, 2.9, and 2.10 hold. Let $T \geq 1$ be an integer, $\beta \in (0, 1]$, and $\varepsilon, \delta \in (0, 1)$ satisfy $d \overset{def}{=} 1 - \eta_1 - \varepsilon^2 > 0$. Use fresh Gaussian or Rademacher JL matrices in Algorithm 1, and set*

$$\rho = \left(1 + c\sqrt{\frac{n}{k}} + c\sqrt{\frac{\log(8T/\delta)}{k}}\right)^2, \qquad \Delta_0 = f_0(x^0) - f_0^{\inf}, \tag{9}$$

where c is the constant in Lemma 2.8. Then, for any stepsize

$$0 < \gamma \leq \frac{d}{2L\rho(1+\varepsilon)(1+\delta_2)}, \tag{10}$$

after T iterations we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x^t)\|^2 \leq \frac{4M\Delta_0}{\gamma\beta GdT} + \frac{2\eta_0}{d} + \frac{2L\gamma\rho(1+\varepsilon)\delta_1}{d}, \tag{11}$$

with probability at least $1 - \delta$, provided that $k \geq C\varepsilon^{-2}\log(8T/\delta)$ for a sufficiently large universal constant $C > 0$.

The proof is given in Appendix D. Neither convexity nor a small upper bound on δ_2 is required Molodtsov et al. (2025). The reference error contributes an explicit residual term, heterogeneity affects both the admissible stepsize and the last term in equation 11, and the factor $M/(\beta G)$ reflects the fraction of honest workers and the weight update. With $S_t = I_n$ ($\varepsilon = 0$, $\rho = 1$), the same analysis gives the uncompressed guarantee.

3 Efficient Implementation

Beyond communication, scalability in large-scale distributed learning is limited by the cost of storing and applying the compression operator: a dense Gaussian or Rademacher matrix $S \in \mathbb{R}^{k\times n}$ requires

$O(nk)$ memory and $O(nk)$ operations per application, which is prohibitive for modern models. We therefore discuss structured and sparse JL constructions, whose geometric guarantees depend on the construction and its parameters, and their use within `Grad-BANT`.

3.1 Efficient JL Transforms

Dense JL matrices can be replaced by the SRHT $S_{\mathrm{SRHT}} = \sqrt{n/k}\, R_T H D$, with a random diagonal sign matrix D, the normalized Walsh–Hadamard matrix H, and a sampler R_T of k coordinates, or by the related FJLT (Ailon & Chazelle, 2006; Tropp, 2011; Ailon & Liberty, 2013), both applied in $O(n \log n)$ time, or by sparse sketches such as CountSketch (Charikar et al., 2002; Clarkson & Woodruff, 2013; Kane & Nelson, 2014), applied in $O(n)$ time. All of them are generated from a random seed, so the server broadcasts only the seed of S_t. The appendix verifies unbiasedness and second-moment bounds for these sketches, which suffices for the expectation analysis (Theorem D.5); the high-probability guarantee of Theorem 2.11 additionally requires the inner-product, Frobenius-norm, and operator-norm events of Appendix D.

3.2 Per-layer Sketching

Applying a single JL transform to the entire parameter vector may be impractical for large models, and even infeasible for large language models. A natural remedy is per-layer sketching, where each layer $x_w \in \mathbb{R}^{n_w}$ of $x = (x_1, \dots, x_W)$ is compressed by its own JL transform $S_w \in \mathbb{R}^{k_w \times n_w}$, i.e., $S = \mathrm{blockdiag}(S_1, \dots, S_W)$. If each layerwise transform preserves norms, so does the full block-diagonal transform.

Theorem 3.1 (Per-layer norm preservation). *Let $x = (x_1, \dots, x_W) \in \mathbb{R}^n \setminus \{0\}$, where $x_w \in \mathbb{R}^{n_w}$ and $\sum_{w=1}^W n_w = n$. Let $S = \mathrm{blockdiag}(S_1, \dots, S_W)$, where $S_w \in \mathbb{R}^{k_w \times n_w}$ are random matrices defined on a common probability space. Suppose that, for each w, there exist parameters $\varepsilon_w \geq 0$ and $\delta_w \in [0,1]$ such that, for every fixed $z \in \mathbb{R}^{n_w}$, $\mathbb{P}\left(\left|\|S_w z\|^2 - \|z\|^2\right| > \varepsilon_w \|z\|^2\right) \leq \delta_w$. Then, for every fixed $x \in \mathbb{R}^n \setminus \{0\}$,*

$$\mathbb{P}\left(\left|\|Sx\|^2 - \|x\|^2\right| > \left(\sum_{w=1}^W \alpha_w(x)\varepsilon_w\right)\|x\|^2\right) \leq \sum_{w=1}^W \delta_w,$$

where $\alpha_w(x) \overset{def}{=} \|x_w\|^2/\|x\|^2$.

If the layerwise distribution of gradient energy is known, the sketch dimensions can be allocated non-uniformly across layers.

Corollary 3.2 (Energy-aware allocation). *Let $\varepsilon, \delta \in (0,1)$ and use Gaussian or Rademacher sketches on the active layers, with dimensions $k_w = \left\lceil C \frac{\log(2W/\delta)}{\varepsilon_w^2} \right\rceil$, where $C > 0$ is a sufficiently large universal constant. Fix $x \in \mathbb{R}^n \setminus \{0\}$ and let $\alpha_w(x) = \|x_w\|^2/\|x\|^2$. For all active layers with $\alpha_w(x) > 0$, set*

$$\varepsilon_w = \frac{\varepsilon\, \alpha_w(x)^{-1/3}}{\sum_{j:\alpha_j(x)>0} \alpha_j(x)^{2/3}}, \qquad \delta_w = \frac{\delta}{W}. \tag{12}$$

Assume that $\varepsilon_w < 1$ and $k_w \leq n_w$ for every active layer. Inactive layers are omitted, with $k_w = 0$. Then, for this fixed vector x, the block-diagonal sketch satisfies an ε-distortion guarantee with probability at least $1 - \delta$, and the total number of transmitted coordinates scales as

$$\sum_{w=1}^W k_w = \Theta\left(\frac{\log(2W/\delta)}{\varepsilon^2}\left(\sum_{w=1}^W \alpha_w(x)^{2/3}\right)^3\right).$$

In this feasible regime, energy-aware allocation can reduce the total communication cost for the fixed vector compared to a uniform allocation.

3.3 Matrix and Tensor Extensions

Neural network parameters are naturally matrices or higher-order tensors, which suggests sketches that respect this structure. For $X \in \mathbb{R}^{n_1 \times n_2}$, instead of applying a dense sketch to $\mathrm{vec}(X)$, one can

use two smaller JL matrices $S_1 \in \mathbb{R}^{k_1 \times n_1}$ and $S_2 \in \mathbb{R}^{k_2 \times n_2}$ and set

$$\mathcal{S}(X) = S_1 X S_2^\top \in \mathbb{R}^{k_1 \times k_2}, \qquad \operatorname{vec}(\mathcal{S}(X)) = (S_2 \otimes S_1)\operatorname{vec}(X).$$

A dense sketch of $\operatorname{vec}(X) \in \mathbb{R}^{n_1 n_2}$ to dimension k requires $O(n_1 n_2 k)$ memory, whereas the Kronecker-structured sketch stores only S_1 and S_2, i.e., $O(n_1 k_1 + n_2 k_2)$ memory, for a compressed dimension $k_1 k_2$. This dimension may not be optimal, but the savings in memory and implementation cost are essential for large-scale models. The same idea extends to higher-order tensors.

Theorem 3.3 (Kronecker-structured JL transform)**.** *Let $S = S_d \otimes S_{d-1} \otimes \ldots \otimes S_1$, where $S_j \in \mathbb{R}^{k_j \times n_j}$ are independent Gaussian or Rademacher JL matrices, and suppose that $\min_{j\in[d]} k_j = \Omega\left(\frac{d^2 \log(2d/\delta)}{\varepsilon^2}\right)$. Then, for every fixed tensor $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, with $\mathcal{S}(X)$ defined by $\operatorname{vec}(\mathcal{S}(X)) = S \operatorname{vec}(X)$,*

$$\mathbb{P}\left(\left|\|\mathcal{S}(X)\|_F^2 - \|X\|_F^2\right| > \varepsilon \|X\|_F^2\right) \le \delta.$$

As for per-layer sketches, their use in the high-probability convergence proof requires the theoretical justifying specified in Appendix D. Overall, dense Gaussian and Rademacher projections are useful for theory and small-scale experiments, whereas structured, sparse, per-layer, and Kronecker-based sketches reduce the cost of applying and storing the compression operator.

4 Experiments

4.1 Comparison of Compressors

In this section we compare several compressors along two axes: the time and memory required to compress and decompress the transmitted parameters and the extent to which they satisfy the JL-properties required by our Algorithm 1.

Table 1 reports the cost of applying each compressor to the parameters of a ResNet18 model. For RandK we treat all parameters as a single flattened vector, whereas for the JL-based compressors and PowerSGD we compress each layer separately. For SRFT and SRHT we consider both scenarios, since these methods can operate in either one without incurring a significant memory or time overhead compared to the remaining JL compressors. CountSketch is omitted from the experiments, as it is not readily supported by standard PyTorch tensor operations. *Stored* denotes the additional memory required to keep the compressor's internal state (e.g. the JL matrix). All measurements are taken at a compression rate of 50%.

Table 1: Cost of applying each compressor to ResNet-18 parameters at $k/n = 0.5$: compression and decompression time, peak memory, memory for the compressor state (Stored) and per-client uplink message size. JL compressors and PowerSGD are applied layer-wise, Rand-k to the flattened parameter vector; † applied to the flattened parameter vector.

Compressor	GPU, ms	CPU, ms	Peak, MB	Stored, MB	Uplink, MB
Gaussian JL, Kronecker (ours)	6.5	**155**	**92.5**	69.9	21.3
Rademacher JL, Kronecker (ours)	8.1	160	92.5	69.9	21.3
SRHT	55.0	718	181.2	117.9	21.3
SRFT	12.7	659	131.1	63.9	21.3
SRHT, whole vector†	30.6	927	644.3	106.6	21.3
SRFT, whole vector†	8.8	693	432.0	64.0	21.3
Rand-k	**3.8**	374	603.4	0.0	21.3
PowerSGD (rank 230)	174.6	195	94.7	18.7	21.3
Gaussian JL†, dense‡	$12 \cdot 10^3$	–	$3.7 \cdot 10^7$	$3.7 \cdot 10^7$	–

As the table shows, our JL variant is the fastest among all JL-based methods, and is outperformed only by RandK on GPU, which is unsurprising since that method merely selects a subset of coordinates. At the same time, in terms of peak memory and CPU time our method outperforms every compressor considered. Applying SRHT and SRFT to the full parameter vector offers no practical advantage over layer-wise application. The effect on running time is marginal (a minor speedup on

GPU and a 5–30% slowdown on CPU), whereas peak memory consumption increases by a factor of 3.3–3.6.

Table 2 compares several quantities that measure how well the JL properties are preserved by the different compression methods. Here Δ_{ip} and Δ_θ denote the quantities

$$\Delta_{\mathrm{ip}} = \frac{\big|\langle Su, Sv\rangle - \langle u, v\rangle\big|}{\|u\|\,\|v\|}, \qquad \Delta_\theta = \big|\angle(Su, Sv) - \angle(u, v)\big|.$$

Δ_{ip} measures the deviation of the inner product of the compressed and uncompressed vectors, whereas the Δ_θ measures the distortion of the angle between the vectors caused by compression.

Table 2: Distortion of inner products (Δ_{ip}) and angles (Δ_θ) under compression, mean $\pm$ std. (a) Random 4096×4096; (b) stochastic ResNet-18 gradient vs. trust-set gradient. All compressors use $k/n = 0.5$; for Adaptive JL, k/n is set implicitly by the target ε from equation 12.

		(a) random matrix 4096×4096 *20 draws of* (S, u, v)*, mean* $\pm$ *std*			*(b) real ResNet-18 gradients 100 pairs, mean* $\pm$ *std*		
Compressor	Variant	k/n	Δ_{ip}, 10^{-4}	Δ_θ, deg	k/n	Δ_{ip}, 10^{-4}	Δ_θ, deg
Kronecker JL (ours)	Rad.	0.500	5.0 ± 4.5	0.011 ± 0.0090	0.500	194 ± 156	1.1 ± 1.7
	Gauss.	0.500	8.7 ± 8.3	0.017 ± 0.013	0.500	94 ± 86	0.8 ± 0.7
SRHT	layer	0.500	2.5 ± 1.9	0.0060 ± 0.0056	0.500	103 ± 80	0.55 ± 0.43
	whole	–	–	–	0.500	302 ± 251	1.31 ± 1.23
SRFT	layer	0.500	2.6 ± 1.7	0.0067 ± 0.0064	0.500	120 ± 125	0.63 ± 0.58
	whole	–	–	–	0.500	470 ± 198	2.1 ± 1.11
Rand-k		0.500	2.1 ± 2.0	0.0063 ± 0.0064	0.500	2000 ± 800	1.4 ± 1.1
Adaptive JL	target $\varepsilon = 0.05$	1.00	0	0	1.00	0	0
	target $\varepsilon = 0.1$	0.841	3.3 ± 2.5	0.011 ± 0.012	1.00	0	0
	target $\varepsilon = 0.2$	0.0610	13 ± 11	0.025 ± 0.021	0.605	39 ± 38	0.23 ± 0.21
	target $\varepsilon = 0.3$	0.0141	19 ± 12	0.057 ± 0.050	0.240	96 ± 77	0.51 ± 0.43

The table reports results for two setups: setup (a) is the compression of two 4096×4096 matrices, and setup (b) is the compression of an actual stochastic gradient of a ResNet18 model and its comparison against the gradient computed on the trust set, as used in Algorithm 1. For all compressors except `AdaptiveJL` the compression rate is fixed at 50%. For `AdaptiveJL` the compression rate is determined implicitly by the parameter ε. We therefore report results for several values of ε. A comparison of the compressors at other compression rates is provided in Appendix A.

As the table shows, RandK performs reasonably well in the first setup, which can be attributed to the random construction of the matrices and to the unbiasedness of the compressor. However, on real gradients the method distorts both the inner product and the angle quite severely. In contrast, the JL-based methods perform well in both scenarios. We also note separately that SRFT and SRHT applied to the gradient as a single vector yield the worst results among all JL-based methods considered, which together with Table 1, indicates that layer-wise compression is preferable.

4.2 Comparison with Byzantine Baselines

This section presents experiments comparing Algorithm 1 against a range of federated learning methods under various Byzantine attack scenarios. All results are reported in the tables below. Every experiment is run in the setting of training a ResNet18 model on the CIFAR-10 dataset. All methods are trained for $T = 200$ communication rounds with the data distributed uniformly across $M = 10$ clients. To compare methods with different compression levels, we measure communication in *full gradient uploads*: a compressed message of k coordinates counts as k/n of a full gradient, where n is the number of model parameters, and uploads are summed over all M clients. After t rounds this gives

$$G(t) = \sum_{s=1}^{t} \sum_{i=1}^{M} \frac{k_{s,i}}{n}, \tag{13}$$

which reduces to $G(t) = tM\,k/n$ for a fixed compression ratio and to $G(t) = tM$ without compression. Over the full $T = 200$ rounds an uncompressed method thus performs 2000 full-gradient

uploads, and a method with $k/n = 0.5$ performs 1000. Details on the attack scenarios and additional experiments are deferred to Appendix A.

Table 3: FedAvg and Grad-BANT with different compressors: (a) no attack, (b) label flipping with 60% Byzantine clients.

Method	Compressor	k/n	Byz.	G$\rightarrow$ 60	G$\rightarrow$ 80	Best acc
(a) no attack						
FedAvg	Gaussian JL	0.1	0	23	52	0.880
FedAvg	Rand-k	0.1	0	15	39	0.897
FedAvg	SRHT	0.1	0	17	42	0.895
Grad-BANT	Gaussian JL	0.1	0	18	53	0.896
Grad-BANT	Rand-k	0.1	0	15	41	0.891
Grad-BANT	SRHT	0.1	0	16	42	0.893
(b) label flipping						
FedAvg	none	1.0	60%	–	–	0.276
FedAvg	Gaussian JL	0.1	60%	–	–	0.247
FedAvg	Rand-k	0.1	60%	–	–	0.315
FedAvg	SRHT	0.1	60%	–	–	0.334
Grad-BANT	Gaussian JL	0.1	60%	31	71	0.838
Grad-BANT	Adaptive JL	0.24	60%	25	69	0.833
Grad-BANT	SRHT	0.1	60%	26	78	0.830
Grad-BANT	Rand-k	0.1	60%	–	–	0.546

Table 4: Best test accuracy of Byzantine-robust methods with compression under label flipping (LF) and random gradients (RG) with 40% and 60% Byzantine clients. All methods use the $k/n = 0.5$.

Method	LF 40%	LF 60%	RG 40%	RG 60%
RoSDHB	0.595	0.331	0.308	0.103
Byz-EF21	0.430	0.294	0.441	0.123
Byz-EF21-BC	0.429	0.265	0.431	0.103
Byz-VR-MARINA 2.0	0.555	0.159	0.571	0.123
Byz-DASHA-PAGE	0.565	0.152	0.455	0.115
Grad-BANT + JL (ours)	**0.883**	**0.862**	**0.620**	0.534

Table 3 reports the results of applying different compressors to `FedAvg` and `GradBant`, both in the attack-free setting and under the LABEL FLIPPING (LF) attack. For each configuration, the table reports the number of full-gradient transmissions required to reach a given value of the target metric, together with the best test accuracy attained. Columns G$\rightarrow$60 and G$\rightarrow$80 report $G(t)$ from equation 13 at the first round t at which the test accuracy reaches 60% and 80%, respectively. For `AdaptiveJL`, k/n denotes the average over rounds and clients.

In the attack-free setting, all compressors perform similarly for both algorithms. Once the attack is enabled, every compressor in the `FedAvg` setting fails to learn and shows almost no accuracy improvement throughout training. In contrast, for `GradBant` the properties of the compressor start to matter. All JL compressors which preserve inner products continue to work, whereas the `RandK` compressor is likewise compromised by the attack.

Table 4 compares Byzantine-robust methods with compression Rammal et al. (2024) under two different attacks. Unlike ours, the remaining methods do not use a trust set, and consequently cannot tolerate a Byzantine majority. When the number of Byzantine clients approaches one half, these methods yield results that lag far behind our algorithm.

Finally, Table 5 compares (a) the best test accuracy and (b) the number of full-gradient uploads G required to reach 60% accuracy for Byzantine-robust baselines without compression and for our algorithm with a compression rate of 50% under several attacks. The results show that GRADBANT requires fewer full-gradient transmissions than the remaining baselines to reach 60% accuracy. We note that under the ALIE attack our method does not reach this level of accuracy, plateauing at 53.9%. In terms of best accuracy, our method outperforms all baselines under the IPM attack. In the attack-free scenario and under SIGNFLIP and LABELFLIP our method stays close to the strongest baselines.

4.3 HETEROGENIOUS SCENARIO

In this section, we consider a scenario with heterogeneous data distribution across clients. To model heterogeneity, we use a Dirichlet(α) distribution to partition the dataset samples among clients. This makes the clients heterogeneous both in local dataset size and in the label distribution of their samples. Table 6 reports the training results of our algorithm under four heterogeneity scenarios, both without any attack and under three different attacks.

All results are reported at a 50% compression rate. As expected, the final model quality degrades as the degree of heterogeneity increases. However, our algorithm remains highly robust to it. In all scenarios, it loses at most 19% accuracy in the worst case and 7% on average relative to the uniform (IID) data distribution.

Table 5: Comparison with Byzantine-robust baselines under different attacks (Byzantine fraction in the header): (a) best test accuracy; (b) full-gradient uploads G to reach 60% test accuracy. Baselines are uncompressed; Grad-BANT + JL uses $k/n = 0.5$. –: not reached.

Method	none	LF 60%	SF 60%	RG 40%	ALIE 60%	IPM 70%
(a) best accuracy						
FedAvg (Adam)	0.917	0.276	0.331	0.132	0.145	0.100
CC	**0.928**	0.828	0.574	0.122	0.143	0.100
CC + bucketing	0.840	**0.897**	**0.891**	0.134	0.143	0.100
CC + FBM	0.924	0.855	**0.892**	0.116	0.184	0.100
Safeguard	0.921	0.104	0.858	0.132	0.113	0.137
FLTrust	0.767	0.704	0.594	0.100	0.100	0.101
RECESS	0.889	0.681	0.512	0.142	0.146	0.106
Zeno	0.919	0.361	0.859	0.105	0.101	0.124
TS-finetuned	0.920	0.884	0.877	0.889	**0.903**	0.630
TS-minimum-grad	0.922	0.887	0.861	**0.901**	0.882	0.677
TS-momentum	0.868	0.863	0.860	0.857	0.865	0.700
Grad-BANT + JL (ours), 50% compression	0.910	0.862	0.878	0.620	0.539	**0.827**
(b) full-gradient uploads G to 60%						
FedAvg (Adam)	160	–	850	–	–	–
CC	**60**	290	–	–	–	–
CC + bucketing	460	360	130	–	–	–
CC + FBM	**60**	140	130	–	–	–
Safeguard	140	–	170	–	–	–
FLTrust	840	1140	–	–	–	–
RECESS	150	610	–	–	–	–
Zeno	160	–	220	–	–	–
TS-finetuned	130	250	170	280	160	1160
TS-minimum-grad	160	160	200	160	**150**	570
TS-momentum	330	330	360	560	310	570
Grad-BANT + JL (ours), full-gradient uploads	80	**120**	**120**	**140**	–	**110**

Table 6: Best test accuracy of Grad-BANT + JL ($k/n = 0.5$) under heterogeneous data split by a Dirichlet distribution with concentration α (smaller α means stronger heterogeneity). Byzantine fractions: 60% for label and sign flipping, 70% for IPM.

Attack	IID	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.2$
No attack	0.910	0.896	0.872	0.842
Label flipping	0.862	0.857	0.782	0.742
Sign flipping	0.878	0.871	0.832	0.753
IPM	0.827	0.826	0.699	0.669

5 Conclusion

We studied Johnson–Lindenstrauss transforms as communication-efficient compressors for Byzantine-robust distributed optimization. Unlike standard sparsification, JL transforms preserve Euclidean geometry with high probability, which suits methods that weight workers by gradient alignment. For a JL-compressed variant of `Grad-BANT`, we proved a high-probability convergence guarantee under smoothness, average data similarity, and an explicit accuracy condition on the trusted reference gradient, without convexity or a small relative heterogeneity parameter.

We also studied structured, sparse, per-layer, and Kronecker-structured sketches, which reduce the cost of storing and applying dense projections in large models; their moment bounds support the expectation analysis, while the high-probability guarantee additionally requires geometric concentration bounds. The experiments confirm that JL-based compression reduces communication while maintaining Byzantine robustness, suggesting that preserving geometric structure is an important design principle for compression in robust distributed learning.

AI USE STATEMENT

In this work, we used generative AI tools for text editing, literature search and code generating. We have also used generative AI to rewrite the proofs to maintain a logical exposition and checking them for arithmetic mistakes. We have not used generative AI tools for deriving the convergence proofs, proposing the new methods or coming up with the ideas. We have reviewed all AI-assisted work, including the results that already have been proven for JL Transforms. As well we checked the generated code We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jayadev Acharya, Chris De Sa, Dylan Foster, and Karthik Sridharan. Distributed learning with sublinear communication. In *International Conference on Machine Learning*, pp. 40–50. PMLR, 2019.

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 557–563, 2006.

Nir Ailon and Edo Liberty. An almost optimal unrestricted fast johnson-lindenstrauss transform. *ACM Transactions on Algorithms*, 9(3):1–12, 2013.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *Advances in neural information processing systems*, 28, 2015.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. *Advances in neural information processing systems*, 23, 2010.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pp. 693–703. Springer, 2002.

Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pp. 81–90, 2013.

Xinyan Dai, Xiao Yan, Kaiwen Zhou, Han Yang, Kelvin KW Ng, James Cheng, and Yu Fan. Hyper-sphere quantization: Communication-efficient sgd for federated learning. *arXiv preprint arXiv:1911.04655*, 2019.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pp. 634–644. IEEE, 1999.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Xiao Guo, Yixuan Qiu, Hai Zhang, and Xiangyu Chang. Randomized spectral co-clustering for large-scale directed networks. *arXiv preprint arXiv:2004.12164*, 2020.

Reinhard Heckel, Michael Tschannen, and Helmut Bölcskei. Dimensionality-reduced subspace clustering. *Information and Inference: A Journal of the IMA*, 6(3):246–283, 2017.

Christina Heinze, Brian McWilliams, and Nicolai Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In *Artificial intelligence and statistics*, pp. 875–883. PMLR, 2016.

William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM*, 61(1):1–23, 2014.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International conference on machine learning*, pp. 5311–5319. PMLR, 2021.

Jon M Kleinberg. Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pp. 599–608, 1997.

Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 614–623, 1998.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Konstantin Mishchenko, Eduard Gorbunov, Martin Takáč, and Peter Richtárik. Distributed learning with compressed gradient differences. *Optimization Methods and Software*, pp. 1–16, 2024.

Gleb Molodtsov, Daniil Medyakov, Sergey Skorik, Nikolas Khachaturov, Shahane Tigranyan, Vladimir Aletov, Aram Avetisyan, Martin Takáč, and Aleksandr Beznosikov. Trial and trust: Addressing byzantine attacks with comprehensive defense strategy. *arXiv preprint arXiv:2505.07614*, 2025.

Cameron Musco and Christopher Musco. Projection-cost-preserving sketches: Proof strategies and constructions. *arXiv preprint arXiv:2004.08434*, 2020.

Ahmad Rammal, Kaja Gruntkowska, Nikita Fedin, Eduard Gorbunov, and Peter Richtárik. Communication compression for byzantine robust learning: New efficient algorithms and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pp. 1207–1215. PMLR, 2024.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *Advances in Neural Information Processing Systems*, 34: 4384–4396, 2021.

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pp. 143–152. IEEE, 2006.

Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pp. 3105–3114. PMLR, 2017.

Mayank Shrivastava, Berivan Isik, Qiaobo Li, Sanmi Koyejo, and Arindam Banerjee. Sketching for distributed deep learning: A sharper analysis. *Advances in Neural Information Processing Systems*, 37:6417–6447, 2024.

Zhao Song, Zheng Yu, and Lichen Zhang. Iterative sketching and its application to federated learning. 2021.

Bin Tang, Michael Shepherd, Malcolm I Heywood, and Xiao Luo. Comparing dimension reduction techniques for document clustering. In *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005, Victoria, Canada, May 9-11, 2005. Proceedings 18*, pp. 292–296. Springer, 2005.

Zhenheng Tang, Shaohuai Shi, Wei Wang, Bo Li, and Xiaowen Chu. Communication-efficient distributed deep learning: A comprehensive survey. *arXiv preprint arXiv:2003.06307*, 2020.

Gian-Andrea Thanei, Christina Heinze, and Nicolai Meinshausen. Random projections for large-scale regression. *Big and Complex Data Analysis: Methodologies and Applications*, pp. 51–68, 2017.

Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Jalaj Upadhyay. Randomness efficient fast-johnson-lindenstrauss transform with applications in differential privacy and compressed sensing. *arXiv preprint arXiv:1410.2470*, 2014.

Thierry Urruty, Chabane Djeraba, and Dan A Simovici. Clustering by random projections. In *Advances in Data Mining. Theoretical Aspects and Applications: 7th Industrial Conference, ICDM 2007, Leipzig, Germany, July 14-18, 2007. Proceedings 7*, pp. 107–119. Springer, 2007.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Haonan Yan, Wenjing Zhang, Qian Chen, Xiaoguang Li, Wenhai Sun, Hui Li, and Xiaodong Lin. Recess vaccine for federated learning: Proactive defense against model poisoning attacks. *Advances in Neural Information Processing Systems*, 36:8702–8713, 2023.

Yue Yu, Jiaxiang Wu, and Longbo Huang. Double quantization for communication-efficient distributed optimization. *Advances in neural information processing systems*, 32, 2019.

A Experimental Details

A.1 Protocol.

Hardware. All experiments were conducted on a single NVIDIA A100 GPU with 80 GB of memory. Running all experiments required approximately 200 GPU-hours in total.

Hyperparameters. The hyperparameters of each method were tuned via grid search. For each method, all tables report the results obtained with the configuration that achieved the highest validation accuracy.

Table 7: Byzantine attacks used in the experiments. Here g_i denotes the stochastic gradient that an honest client i would send, $\bar{g}$ and σ denote the coordinate-wise mean and standard deviation of the honest clients' gradients, and C is the number of classes.

Attack	Byzantine behavior	Byz. fraction
Label flipping (LF)	Data poisoning: Byzantine clients compute gradients on their local data with each label y replaced by $C-1-y$.	40%, 60%
Sign flipping (SF)	Byzantine clients send the negated honest gradient $-g_i$.	60%
Random gradients (RG)	Byzantine clients send random vectors with i.i.d. Gaussian entries, $\xi \sim \mathcal{N}(0, s^2 I)$.	40%, 60%
ALIE	Omniscient attack: Byzantine clients send $\bar{g} - z\,\sigma$, a small coordinate-wise perturbation of the honest mean that stays within the spread of honest gradients and thus evades robust aggregation rules.	60%
IPM	Omniscient attack: Byzantine clients send $-\epsilon\,\bar{g}$ with $\epsilon > 0$, so that the inner product between the aggregated update and the true gradient becomes negative.	70%

A.2 Trial ablation

In this section, we study how the size and distribution of the trust set affect the performance of Algorithm 1. All experiments use $k/n = 0.5$, label flipping with 60% Byzantine clients, IID data, and 100 training rounds. The results are summarized in Table 8.

Trust set size. Table 8 (a) shows that the size of the trust set has little effect on either the best test accuracy or the convergence speed. As the size increases from 100 to 2000 images, the best accuracy varies only between 0.838 and 0.844, and the number of rounds needed to reach 60% and 80% accuracy changes by at most 6 and 5 rounds, respectively.

Trust set distribution. Table 8 (b) considers a trust set of 1500 images whose distribution differs from uniform. Shifts in image appearance slow down training but do not prevent convergence. With grayscale images, reaching 60% and 80% accuracy takes 6 and 13 more rounds, respectively, while the best accuracy decreases only slightly, from 0.840 to 0.831. Strong augmentation has a larger effect. Reaching 60% accuracy takes almost three times as many rounds, 80% is never reached, and the best accuracy is 0.761. In both cases, training remains stable, and the final accuracy coincides with the best one. In contrast, class coverage is critical. When the trust set contains only 5 of the 10 classes, the best accuracy remains high (0.828), but training becomes unstable and the final accuracy drops to 0.636. With only 2 classes, the method fails completely: the accuracy never exceeds 0.16, which is close to random guessing.

Overall, Algorithm 1 is robust to the size of the trust set and to moderate shifts in its appearance. The key requirement is that the trust set covers all classes; its size mainly affects the stability of training in later rounds.

Training on the trust set alone. To check whether federated training is necessary when a trust set is available, we train ResNet-18 centrally on the trust set alone, using the same optimizer (Adam,

Table 8: Effect of the trust set on Grad-BANT + JL: (a) size; (b) distribution shift, 1500 images. R→ 60 and R→ 80: rounds to reach 60% and 80% test accuracy; Best and Final: best and last-round accuracy; –: not reached. Label flipping, 60% Byzantine clients, $k/n = 0.5$, IID data, 100 rounds.

Trust set	R→ 60	R→ 80	Best	Final
(a) size, all 10 *classes*				
100	**21**	**61**	**0.844**	0.683
500	**21**	65	0.841	0.674
1000	23	64	0.839	0.622
1500	22	65	0.840	0.835
2000	27	66	0.838	**0.838**
(b) shift, 1500 *images*				
none	**22**	**65**	**0.840**	**0.835**
grayscale	28	78	0.831	0.831
strong augmentation	63	–	0.761	0.761
5 of 10 classes	23	71	0.828	0.636
2 of 10 classes	–	–	0.160	0.102

Table 9: Training ResNet-18 on the trust set alone, without federation. @r: CIFAR-10 test accuracy after $62r$ optimizer steps, i.e., the computation of one client over r rounds. Epochs: passes over the trust set. Mean ± std over 3 seeds; each seed samples a different trust set. Federated training with the same trust set sizes reaches 0.89–0.92 (Table 8 (a)).

Trust set	Epochs	@20	@50	@100	@200	Best
100	7936	0.258 ± 0.019	0.260 ± 0.014	0.280 ± 0.012	0.270 ± 0.015	0.287 ± 0.010
500	1587	0.387 ± 0.021	0.407 ± 0.027	0.410 ± 0.046	0.428 ± 0.024	0.439 ± 0.023
1000	794	0.466 ± 0.015	0.496 ± 0.011	0.515 ± 0.016	0.501 ± 0.003	0.542 ± 0.012
1500	529	0.523 ± 0.017	0.554 ± 0.021	0.577 ± 0.015	0.578 ± 0.008	0.609 ± 0.006
2000	397	0.554 ± 0.015	0.607 ± 0.012	0.615 ± 0.012	0.628 ± 0.011	0.653 ± 0.004

learning rate 0.003), batch size (64), and augmentation as the clients. The budget is 12 400 optimizer steps, which matches the local computation of one client over 200 rounds. Table 9 shows that this baseline is far behind federated training. With 1500 images, the best accuracy is 0.609, whereas Algorithm 1 with a trust set of the same size reaches 0.913 within 100 rounds, even with 60% Byzantine clients (Table 8). The gap grows as the trust set shrinks: with 100 images, training on the trust set alone reaches at most 0.287, while the federated method still reaches 0.894. Even with 2000 images, the best accuracy is only 0.653. Moreover, the accuracy saturates early: between the 100th and 200th round-equivalents, it changes by less than 0.02 for all sizes, and the final accuracy is below the best one, which suggests overfitting to the small sample.

A.3 Additional experiments

Table 10 extends the comparison of Table 2 to compression rates $k/n \in \{0.5, 0.3, 0.1\}$ in the same setting: the distortion is measured between the stochastic ResNet-18 gradient and the trust-set gradient over 100 pairs.

As expected, the inner-product distortion Δ_{ip} increases monotonically for all compressors as the compression becomes stronger. Reducing k/n from 0.5 to 0.1 increases Δ_{ip} by a factor of 3.0–5.2 for Kronecker JL and of about 5 for layer-wise SRHT and SRFT.

Kronecker JL remains competitive with layer-wise SRHT and SRFT at all compression rates: at every k/n, one of its variants attains the smallest mean Δ_{ip} among all compressors, although the differences are within one standard deviation. The angle errors of all JL-type methods are of the same order, ranging from 0.55° at $k/n = 0.5$ to 4.2° at $k/n = 0.1$.

Applying SRHT and SRFT to the whole parameter vector does not reduce the distortion. At $k/n = 0.5$, it increases Δ_{ip} by a factor of 2.9 for SRHT and 3.9 for SRFT; at smaller k/n, the whole-vector variants are comparable to the layer-wise ones or worse by up to a factor of 1.5.

Rand-k yields the largest Δ_{ip} at every compression rate, 4–38 times larger than that of the other compressors, and its distortion grows the fastest as k/n decreases. At the same time, its angle error remains comparable to that of the other methods.

Finally, Adaptive JL with target $\varepsilon = 0.3$ selects $k/n = 0.240$ and attains Δ_{ip} comparable to that of Gaussian Kronecker JL at $k/n = 0.5$ (96 vs. $94 \cdot 10^{-4}$), i.e., with less than half the message size.

Table 10: Distortion of inner products (Δ_{ip}) and angles (Δ_θ) under compression at different compression rates: stochastic ResNet-18 gradient vs. trust-set gradient, 100 pairs, mean $\pm$ std. The rows for $k/n = 0.5$ and for Adaptive JL are repeated from Table 2. For Adaptive JL, $k/n = 0.240$ is set implicitly by the target $\varepsilon = 0.3$ from equation 12.

Compressor	Variant	Δ_{ip}, 10^{-4}	Δ_θ, deg
$k/n = 0.5$			
Kronecker JL (ours)	Rad.	194 ± 156	1.1 ± 1.7
	Gauss.	94 ± 86	0.8 ± 0.7
SRHT	layer	103 ± 80	0.55 ± 0.43
	whole	302 ± 251	1.31 ± 1.23
SRFT	layer	120 ± 125	0.63 ± 0.58
	whole	470 ± 198	2.1 ± 1.11
Rand-k		2000 ± 800	1.4 ± 1.1
$k/n = 0.3$			
Kronecker JL (ours)	Rad.	$300 \pm 19?$	1.63 ± 0.98
	Gauss.	390 ± 250	1.69 ± 1.35
SRHT	layer	474 ± 283	1.72 ± 0.96
	whole	456 ± 341	2.63 ± 2.11
SRFT	layer	534 ± 313	2.13 ± 2.09
	whole	632 ± 412	1.87 ± 1.31
Rand-k		4963 ± 914	1.57 ± 1.16
$k/n = 0.1$			
Kronecker JL (ours)	Rad.	590 ± 490	3.51 ± 2.93
	Gauss.	490 ± 490	4.17 ± 3.92
SRHT	layer	512 ± 481	2.1 ± 1.4
	whole	715 ± 653	4.15 ± 3.31
SRFT	layer	614 ± 541	3.04 ± 2.98
	whole	891 ± 753	3.1 ± 2.9
Rand-k		18500 ± 7900	3.3 ± 2.4
adaptive k/n			
Adaptive JL	target $\varepsilon = 0.3$	96 ± 77	0.51 ± 0.43

Table 11 shows that stronger compression substantially reduces the communication required to reach 80% test accuracy. Moderate compression does not slow down training: in all settings, the number of rounds needed to reach 80% is smallest at $k/n = 0.5$. At $k/n = 0.1$, more rounds are required, but each round is much cheaper, so the total traffic decreases by a factor of 2.8–4.0 compared to $k/n = 0.5$ and by a factor of 6.7–10.6 compared to $k/n = 0.9$. At the same time, the best accuracy at $k/n = 0.1$ is at most 0.018 lower than at $k/n = 0.5$.

B Technical Statements

Statement below is crucial for deriving high-probability convergence of proposed methods

Lemma B.1. *Suppose we have events $E_1, E_2, \ldots, E_n$ that holds with following probabilities:*

$$\mathbb{P}[E_i] \geq 1 - \delta_i,$$

Then,

$$\mathbb{P}\left[\bigcap_{i=1}^{n} E_i\right] \geq 1 - \sum_{i=1}^{n} \delta_i$$

Table 11: Communication efficiency of Grad-BANT + JL for different compression rates k/n (IID data, 200 rounds, single seed): (a) rounds and (b) traffic in full-gradient uploads (a round costs k/n units) to reach 80% test accuracy; (c) best test accuracy. Columns 0.9–0.5 use Rademacher JL, columns 0.3–0.1 use Gaussian JL.

k/n	0.9	0.7	0.5	0.3	0.1
(a) rounds to 80%					
No attack	54	45	41	45	73
Label flipping 60%	71	62	57	60	71
Sign flipping 60%	68	47	46	61	58
IPM 70%	68	56	48	51	68
(b) traffic to 80%*, full-gradient uploads*					
No attack	48.6	31.5	20.5	13.5	**7.3**
Label flipping 60%	63.9	43.4	28.5	18.0	**7.1**
Sign flipping 60%	61.2	32.9	23.0	18.3	**5.8**
IPM 70%	61.2	39.2	24.0	15.3	**6.8**
(c) best accuracy in 200 *rounds*					
No attack	0.900	0.906	0.910	0.906	0.893
Label flipping 60%	0.853	0.868	0.862	0.865	0.857
Sign flipping 60%	0.857	0.875	0.878	0.868	0.860
IPM 70%	0.854	0.875	0.827	0.870	0.861

Proof.

$$\mathbb{P}\left[\bigcap_{i=1}^{n} E_i\right] = 1 - \mathbb{P}\left[\bigcup_{i=1}^{n} \overline{E_i}\right] \geq 1 - \sum_{i=1}^{n} \mathbb{P}\left[\overline{E_i}\right] \geq 1 - \sum_{i=1}^{n} \delta_i$$

□

Next lemma will be useful throughout some theorems, allowing us to deal with norms:

Lemma B.2 (Young's inequality)**.** *Let $x, y \in \mathbb{R}^n$. Then, for any $\alpha > 0$ we have*

$$\langle x, y\rangle \leq \frac{1}{2\alpha}\|x\|^2 + \frac{\alpha}{2}\|y\|^2.$$

C JL STATEMENTS

We start with the properties of a single sketch: inner-product preservation, moments of the lifted vector, and high-probability norm bounds. We then show how these properties extend to structured matrices, separate layers, and tensor modes. Throughout, Sx is the compressed vector and $S^\top Sx$ is the lifted vector used in the update. We write $\|\cdot\|$ for the Euclidean norm of a vector or the operator norm of a matrix, and $\|\cdot\|_F$ for the Frobenius norm of a matrix or tensor.

C.1 FROM NORM PRESERVATION TO INNER PRODUCTS

Lemma C.1 (Preservation of scalar product)**.** *Let S be a linear JL transform with distortion ε and one-vector failure probability δ. For every fixed $x, y \in \mathbb{R}^n$,*

$$\mathbb{P}\left[|\langle Sx, Sy\rangle - \langle x, y\rangle| \le \varepsilon \|x\|\|y\|\right] \ge 1 - 2\delta.$$

Proof. The claim is immediate if either vector is zero. By linearity, it is enough to prove it for unit vectors. Apply the JL property to $x + y$ and $x - y$. With probability at least $1 - 2\delta$, both squared norms are preserved within relative error ε. On this event, polarization and the parallelogram identity give

$$\begin{aligned}|\langle Sx, Sy\rangle - \langle x, y\rangle| &\le \frac{\varepsilon}{4}\left(\|x + y\|^2 + \|x - y\|^2\right) \\ &= \frac{\varepsilon}{2}\left(\|x\|^2 + \|y\|^2\right) = \varepsilon.\end{aligned}$$

For general nonzero vectors, apply this bound to $x/\|x\|$ and $y/\|y\|$, and multiply by $\|x\|\|y\|$. □

Lemma C.2 (Simultaneous upper JL bound)**.** *Let S be a linear JL transform with distortion ε and one-vector failure probability δ. For fixed vectors $v_1, \dots, v_m \in \mathbb{R}^n$,*

$$\mathbb{P}\left[\|Sv_i\|^2 \le (1 + \varepsilon)\|v_i\|^2 \quad \text{for all } i = 1, \dots, m\right] \ge 1 - m\delta.$$

Proof. Each upper bound fails with probability at most δ. Take the union bound over the m vectors. □

C.2 DENSE SKETCHES: MOMENTS AND OPERATOR NORMS

The next identity gives the moments of both dense sketches and the sparse Gaussian construction below. It also allows an independent orthogonal transformation before sketching.

Lemma C.3 (Moments of a normalized i.i.d. matrix)**.** *Let $A \in \mathbb{R}^{k\times n}$ have i.i.d. entries with*

$$\mathbb{E}A_{11} = 0, \qquad \mathbb{E}A_{11}^2 = 1, \qquad \mathbb{E}A_{11}^4 = \mu_4 < \infty.$$

Let $U \in \mathbb{R}^{n\times n}$ be orthogonal and independent of A; U may be deterministic or random. For $S = AU/\sqrt{k}$,

$$\mathbb{E}[S^\top S] = I_n, \qquad \mathbb{E}\|S^\top Sx\|^2 = \left(1 + \frac{n - 2 + \mu_4}{k}\right)\|x\|^2.$$

The second identity holds for every fixed $x \in \mathbb{R}^n$.

Proof. First take $U = I_n$, and write the rows of A as column vectors $a_1, \dots, a_k$. Since $\mathbb{E}[a_r a_r^\top] = I_n$,

$$\mathbb{E}[S^\top S] = \frac{1}{k}\sum_{r=1}^{k}\mathbb{E}[a_r a_r^\top] = I_n.$$

For the second moment, expand

$$\mathbb{E}\|A^\top Ax\|^2 = \sum_{r,s=1}^{k}\mathbb{E}\left[\langle a_r, a_s\rangle\langle a_r, x\rangle\langle a_s, x\rangle\right].$$

If $r \neq s$, conditioning on a_r shows that the corresponding expectation is $\mathbb{E}\langle a_r, x\rangle^2 = \|x\|^2$. For a diagonal term, independence and centering of the entries remove all mixed terms, giving

$$\mathbb{E}\left[\|a_r\|^2\langle a_r, x\rangle^2\right] = \sum_{j=1}^{n}\left(\sum_{i\neq j} x_i^2 + \mu_4 x_j^2\right) = (n-1+\mu_4)\|x\|^2.$$

There are $k(k-1)$ off-diagonal terms and k diagonal terms. Dividing their sum by k^2 yields

$$\mathbb{E}\|S^\top Sx\|^2 = \frac{k(k-1)+k(n-1+\mu_4)}{k^2}\|x\|^2 = \left(1+\frac{n-2+\mu_4}{k}\right)\|x\|^2.$$

For general U, condition on U. Orthogonality reduces the calculation to the case just proved with x replaced by Ux, whose norm equals $\|x\|$. Averaging over U proves both identities. □

Lemma C.4 (Gaussian and Rademacher moments)**.** *Let $S \in \mathbb{R}^{k\times n}$ have independent entries, either $S_{ij} \sim \mathcal{N}(0, 1/k)$ or $S_{ij} \in \{-1/\sqrt{k}, 1/\sqrt{k}\}$ with equal probabilities. In both cases $\mathbb{E}[S^\top S] = I_n$, and for every fixed $x \in \mathbb{R}^n$,*

$$\mathbb{E}\|S^\top Sx\|^2 = \begin{cases} \dfrac{n+k+1}{k}\|x\|^2, & \textit{Gaussian}, \\ \dfrac{n+k-1}{k}\|x\|^2, & \textit{Rademacher}. \end{cases}$$

Proof. Apply Lemma C.3 with $U = I_n$. The fourth moment of a standard Gaussian entry is 3, and that of a Rademacher entry is 1. □

We next bound the operator norm. Every such bound also controls the reconstructed vector, since for every matrix S and every x,

$$\|S^\top S\| = \|S\|^2, \qquad \|S^\top Sx\|^2 \le \|S\|^4\|x\|^2.$$

The following standard estimate will be used for the Rademacher and sparse Gaussian constructions. Its subgaussian norm is defined, for a scalar X and a random vector Z, by

$$\|X\|_{\psi_2} := \inf\{s > 0 : \mathbb{E}e^{X^2/s^2} \le 2\}, \qquad \|Z\|_{\psi_2} := \sup_{\|u\|=1}\|\langle Z, u\rangle\|_{\psi_2}.$$

Lemma C.5 (Subgaussian operator-norm estimate (Vershynin, 2026))**.** *Let $A \in \mathbb{R}^{m\times n}$ have independent, mean-zero, isotropic, subgaussian rows $A_1, \dots, A_m$, with $\|A_i\|_{\psi_2} \le K$ for all i. There is a universal constant $C > 0$ such that, for every $t \ge 0$,*

$$\mathbb{P}\left[\|A\| \le \sqrt{m} + CK^2(\sqrt{n}+t)\right] \ge 1 - 2e^{-t^2}.$$

Lemma C.6 (Gaussian and Rademacher operator norms)**.** *There is a universal constant $C > 0$ such that, for a Gaussian or a Rademacher JL matrix $S \in \mathbb{R}^{k\times n}$ and every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\|S\| \le 1 + C\left(\sqrt{\frac{n}{k}} + \sqrt{\frac{\log(2/\delta)}{k}}\right).$$

Squaring the right-hand side bounds $\|S^\top S\|$; its fourth power times $\|x\|^2$ bounds $\|S^\top Sx\|^2$ simultaneously for all $x \in \mathbb{R}^n$.

Proof. In the Gaussian case, write $S = A/\sqrt{k}$ with standard Gaussian entries in A. By Vershynin (2026, Corollary 7.3.2), there is a universal constant $c_0 > 0$ such that, for every $t \ge 0$,

$$\mathbb{P}\left[\|A\| \ge \sqrt{k} + \sqrt{n} + t\right] \le 2e^{-c_0 t^2}.$$

Taking $t = \sqrt{\log(2/\delta)/c_0}$ and dividing by $\sqrt{k}$ gives, with probability at least $1 - \delta$,

$$\|S\| \le 1 + \sqrt{\frac{n}{k}} + \frac{1}{\sqrt{c_0}}\sqrt{\frac{\log(2/\delta)}{k}},$$

which is the stated bound for any $C \ge \max\{1, c_0^{-1/2}\}$.

In the Rademacher case, write again $S = A/\sqrt{k}$. Its independent rows are mean-zero and isotropic. For each unit vector u, Hoeffding's lemma gives

$$\mathbb{E}e^{\lambda\langle A_i, u\rangle} \le e^{\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R},$$

so their subgaussian norms are bounded by a universal constant. Apply Lemma C.5 with $m = k$ and $t = \sqrt{\log(2/\delta)}$, and divide by $\sqrt{k}$. Enlarging C if necessary, one constant serves both constructions. The remaining claims follow from $\|S^\top S\| = \|S\|^2$ and $\|S^\top Sx\| \le \|S^\top S\|\|x\|$. □

C.3 Preserving the total squared norm of several vectors

The convergence proof needs a bound on the sum of the squared norms of the compressed honest gradients. Applying a separate JL guarantee to every gradient would introduce a failure-probability factor equal to the number of honest workers. For Gaussian and Rademacher sketches, we can instead control this sum directly.

Place the vectors in the columns of a fixed matrix $V \in \mathbb{R}^{n\times N}$. Then $\|V\|_F^2$ is their total squared norm and $\|SV\|_F^2$ is the corresponding quantity after compression. We first give an elementary concentration bound used in the Gaussian case.

Lemma C.7 (Weighted chi-square concentration)**.** *Let $a_1, \dots, a_r \ge 0$ and let $Z_1, \dots, Z_r$ be independent χ_k^2 random variables. For every $\varepsilon \in (0, 1)$,*

$$\mathbb{P}\left[\left|\frac{1}{k}\sum_{i=1}^r a_i Z_i - \sum_{i=1}^r a_i\right| > \varepsilon \sum_{i=1}^r a_i\right] \le 2e^{-k\varepsilon^2/16}.$$

Proof. The claim is immediate if all weights vanish. Otherwise, divide the weights by their sum, so that $\sum_i a_i = 1$. For a single collection of independent standard Gaussian variables g_i, the Gaussian moment-generating function gives, for $0 \le \lambda \le 1/4$,

$$\log \mathbb{E}\exp\left(\lambda \sum_i a_i(g_i^2 - 1)\right) = -\lambda - \frac{1}{2}\sum_i \log(1 - 2\lambda a_i) \le 4\lambda^2 \sum_i a_i^2 \le 4\lambda^2,$$

$$\log \mathbb{E}\exp\left(-\lambda \sum_i a_i(g_i^2 - 1)\right) = \lambda - \frac{1}{2}\sum_i \log(1 + 2\lambda a_i) \le \lambda^2 \sum_i a_i^2 \le \lambda^2.$$

The second bound also holds for every $\lambda \ge 0$. Write each Z_i as the sum of k independent squared standard Gaussian variables. Independence across these k repetitions and the exponential Markov inequality give upper and lower tail bounds

$$e^{-k\lambda\varepsilon + 4k\lambda^2} \quad \text{and} \quad e^{-k\lambda\varepsilon + k\lambda^2},$$

respectively. Taking $\lambda = \varepsilon/8$ in the first bound and $\lambda = \varepsilon/2$ in the second proves the result. □

Lemma C.8 (Preservation of a fixed Frobenius norm)**.** *Let $S \in \mathbb{R}^{k\times n}$ be a Gaussian or Rademacher JL matrix. There is a universal constant $c_1 > 0$ such that, for every fixed matrix $V \in \mathbb{R}^{n\times N}$ and every $\varepsilon \in (0, 1)$,*

$$\mathbb{P}\left[\left|\|SV\|_F^2 - \|V\|_F^2\right| > \varepsilon\|V\|_F^2\right] \le 2e^{-c_1 k\varepsilon^2}.$$

In particular, the bound does not depend on the number of columns N.

Proof. The case $V = 0$ is immediate.

Gaussian sketches. Take a compact singular value decomposition of V, with nonzero singular values $\sigma_1, \dots, \sigma_r$. Gaussian rotational invariance gives

$$\|SV\|_F^2 \overset{d}{=} \frac{1}{k}\sum_{i=1}^r \sigma_i^2 Z_i, \qquad \|V\|_F^2 = \sum_{i=1}^r \sigma_i^2,$$

where the Z_i are independent χ_k^2 variables. Apply Lemma C.7 with weights σ_i^2.

Rademacher sketches. Let $\xi_1, \dots, \xi_k$ be the rows of $\sqrt{k}S$, viewed as column vectors, and set $B = VV^\top$. Then

$$k\|SV\|_F^2 = \sum_{\ell=1}^k \xi_\ell^\top B \xi_\ell, \qquad \mathbb{E}\|SV\|_F^2 = \operatorname{tr}(B) = \|V\|_F^2.$$

Apply the Hanson–Wright inequality to the concatenated vector of all kn independent signs and the block-diagonal matrix with k copies of B. For every $s > 0$, this gives

$$\mathbb{P}\left[\left|k\|SV\|_F^2 - k\|V\|_F^2\right| > s\right] \le 2\exp\left(-c\min\left\{\frac{s^2}{k\|B\|_F^2}, \frac{s}{\|B\|}\right\}\right).$$

Since B is positive semidefinite, both $\|B\|$ and $\|B\|_F$ are at most $\mathrm{tr}(B) = \|V\|_F^2$. With $s = k\varepsilon\|V\|_F^2$ and $0 < \varepsilon < 1$, the minimum is at least $k\varepsilon^2$. Choosing a common constant for the two constructions completes the proof. □

This estimate concerns one fixed matrix V, rather than all matrices simultaneously. It also applies conditionally when V is determined by the past and S is sampled independently afterwards. Together with an operator-norm bound, it controls the lifted vectors through

$$\|S^\top SV\|_F^2 \le \|S\|^2\|SV\|_F^2.$$

The same Frobenius-norm estimate will allow us to compress a tensor one mode at a time below.

C.4 Structured sketches

The following results give the same moment and operator-norm information for SRHT, the FJLT with a sparse Gaussian core, and CountSketch.

Lemma C.9 (SRHT). *Assume that n is a power of two and $1 \le k \le n$. Let H be the normalized Walsh–Hadamard matrix, let D have independent Rademacher entries on its diagonal, and let $T \subset [n]$ be a uniformly random k-element subset independent of D. If R_T restricts a vector to its coordinates in T, define*

$$S = \sqrt{\frac{n}{k}}\, R_T H D.$$

Then

$$\mathbb{E}[S^\top S] = I_n, \qquad \mathbb{E}\|S^\top Sx\|^2 = \frac{n}{k}\|x\|^2.$$

The second identity holds for every fixed $x \in \mathbb{R}^n$. Moreover, $\|S\| = \sqrt{n/k}$ almost surely.

Proof. The matrix $R_T^\top R_T$ is the coordinate projector onto T. Each coordinate is selected with probability k/n, so

$$\mathbb{E}[R_T^\top R_T] = \frac{k}{n} I_n.$$

Since HD is orthogonal, conditioning on D gives

$$\mathbb{E}[S^\top S \mid D] = \frac{n}{k} D H^\top \mathbb{E}[R_T^\top R_T] HD = I_n.$$

Also, $R_T R_T^\top = I_k$ implies

$$SS^\top = \frac{n}{k} I_k, \qquad (S^\top S)^2 = \frac{n}{k} S^\top S.$$

The first identity gives $\|S\| = \sqrt{n/k}$. The second and unbiasedness give

$$\mathbb{E}\|S^\top Sx\|^2 = x^\top \mathbb{E}[(S^\top S)^2]x = \frac{n}{k}\|x\|^2.$$

□

Lemma C.10 (FJLT with a sparse Gaussian core). *Assume that n is a power of two. Let H be the normalized Walsh–Hadamard matrix, let D be a diagonal Rademacher matrix, and let $P \in \mathbb{R}^{k\times n}$ have independent entries*

$$P_{ij} = \begin{cases} 0, & \text{with probability } 1-q, \\ g_{ij}/\sqrt{q}, & \text{with probability } q, \end{cases} \qquad g_{ij} \sim \mathcal{N}(0,1), \qquad q \in (0,1].$$

Assume that all random choices are independent, and define $S = PHD/\sqrt{k}$. Then

$$\mathbb{E}[S^\top S] = I_n, \qquad \mathbb{E}\|S^\top Sx\|^2 = \left(1 + \frac{n-2+3/q}{k}\right)\|x\|^2 \quad \text{for every fixed } x \in \mathbb{R}^n.$$

Moreover, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|S\| \le 1 + \frac{C}{q}\left(\sqrt{\frac{n}{k}} + \sqrt{\frac{\log(2/\delta)}{k}}\right),$$

where $C > 0$ is universal.

Proof. The entries of P satisfy

$$\mathbb{E}P_{11} = 0, \qquad \mathbb{E}P_{11}^2 = 1, \qquad \mathbb{E}P_{11}^4 = 3/q.$$

The moment identities follow from Lemma C.3 with $A = P$ and the independent orthogonal matrix $U = HD$.

To bound the operator norm, observe that

$$\mathbb{E}e^{\lambda P_{11}} = 1 - q + qe^{\lambda^2/(2q)} \le e^{\lambda^2/(2q)}.$$

By independence, for a row p of P and a unit vector u,

$$\mathbb{E}e^{\lambda\langle p,u\rangle} \le \prod_{j=1}^{n} e^{\lambda^2 u_j^2/(2q)} = e^{\lambda^2/(2q)}.$$

Thus the rows are mean-zero and isotropic, with subgaussian norms at most $C/\sqrt{q}$. Lemma C.5, with $t = \sqrt{\log(2/\delta)}$, gives

$$\|P\| \le \sqrt{k} + \frac{C}{q}\left(\sqrt{n} + \sqrt{\log(2/\delta)}\right)$$

with the claimed probability, after changing the universal constant. Finally, orthogonality of HD gives $\|S\| = \|P\|/\sqrt{k}$. □

Lemma C.11 (CountSketch)**.** *Let $h(1), \dots, h(n)$ be independent and uniform on $[k]$, and let $\sigma(1), \dots, \sigma(n)$ be independent Rademacher signs, independent of h. Define $S \in \mathbb{R}^{k\times n}$ by*

$$S_{rj} = \sigma(j)\mathbf{1}_{\{h(j)=r\}}.$$

Then

$$\mathbb{E}[S^\top S] = I_n, \qquad \mathbb{E}\|S^\top Sx\|^2 = \left(1 + \frac{n-1}{k}\right)\|x\|^2 \quad \textit{for every fixed } x \in \mathbb{R}^n.$$

If $N_r = \#\{j \in [n] : h(j) = r\}$ is the number of coordinates assigned to row r, then

$$SS^\top = \operatorname{diag}(N_1, \dots, N_k), \qquad \|S\|^2 = \max_{1\le r\le k} N_r.$$

Consequently, for every $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|S\|^2 \le \frac{n}{k} + \sqrt{\frac{2n}{k}\log\frac{k}{\delta}} + \frac{2}{3}\log\frac{k}{\delta}.$$

Proof. Every column contains one nonzero entry of magnitude one. Hence $(S^\top S)_{jj} = 1$, while for $i \ne j$,

$$(S^\top S)_{ij} = \sigma(i)\sigma(j)\mathbf{1}_{\{h(i)=h(j)\}}.$$

Independent centered signs make the off-diagonal expectations zero, proving unbiasedness. To compute the second moment, write

$$(S^\top Sx - x)_j = \sigma(j)\sum_{i\ne j}\sigma(i)\mathbf{1}_{\{h(i)=h(j)\}}x_i.$$

After squaring, the mixed terms vanish in expectation. Since each pair of coordinates collides with probability $1/k$,

$$\mathbb{E}\|S^\top Sx - x\|^2 = \frac{1}{k}\sum_{j=1}^{n}\sum_{i\ne j} x_i^2 = \frac{n-1}{k}\|x\|^2.$$

Unbiasedness now gives the claimed second moment.

Distinct rows of S have disjoint supports, and the squared norm of row r is N_r. This proves the formula for $SS^\top$ and its operator norm. For each r, the count N_r is binomial with mean n/k. Bernstein's inequality gives, for every $u \ge 0$,

$$\mathbb{P}\left[N_r > \frac{n}{k} + \sqrt{\frac{2nu}{k}} + \frac{2u}{3}\right] \le e^{-u}.$$

Take $u = \log(k/\delta)$ and apply the union bound over the k rows. □

C.5 Per-layer sketching

We now combine the one-layer guaranties using the block-diagonal structure of the sketch. Write

$$x = (x_1, \dots, x_W), \qquad x_w \in \mathbb{R}^{n_w}, \qquad n = \sum_{w=1}^{W} n_w, \qquad S = \operatorname{diag}(S_1, \dots, S_W),$$

where $S_w \in \mathbb{R}^{k_w \times n_w}$. The key identities are

$$\|x\|^2 = \sum_{w=1}^{W} \|x_w\|^2, \qquad \|Sx\|^2 = \sum_{w=1}^{W} \|S_w x_w\|^2.$$

$$S^\top S = \operatorname{diag}(S_1^\top S_1, \dots, S_W^\top S_W), \qquad \|S^\top S x\|^2 = \sum_{w=1}^{W} \|S_w^\top S_w x_w\|^2.$$

As in the main text, for $x \neq 0$ we use the energy fractions

$$\alpha_w(x) = \frac{\|x_w\|^2}{\|x\|^2}, \qquad \sum_{w=1}^{W} \alpha_w(x) = 1.$$

Thus, squared-norm bounds combine as energy-weighted averages, whereas operator-norm bounds combine by taking the largest layer bound.

Theorem C.12 (Per-layer JL properties)**.** *Suppose that, for each w, there are $\varepsilon_w \geq 0$ and $\delta_w \in [0,1]$ such that, for every fixed $u \in \mathbb{R}^{n_w}$,*

$$\mathbb{P}\left[\left|\|S_w u\|^2 - \|u\|^2\right| > \varepsilon_w \|u\|^2\right] \leq \delta_w.$$

Then, for every fixed $x \neq 0$,

$$\mathbb{P}\left[\left|\|Sx\|^2 - \|x\|^2\right| > \left(\sum_{w=1}^{W} \alpha_w(x)\varepsilon_w\right) \|x\|^2\right] \leq \sum_{w=1}^{W} \delta_w.$$

In particular, if $\max_w \varepsilon_w \leq \varepsilon$, then S satisfies the $(\varepsilon, \sum_w \delta_w)$ distributional JL property. If the blocks are independent, both failure bounds can be replaced by $1 - \prod_{w=1}^{W}(1-\delta_w)$.

Proof. Fix $x \neq 0$. With probability at least $1 - \sum_w \delta_w$, every layer satisfies $|\|S_w x_w\|^2 - \|x_w\|^2| \leq \varepsilon_w \|x_w\|^2$. On this event,

$$\left|\|Sx\|^2 - \|x\|^2\right| \leq \sum_{w=1}^{W} \left|\|S_w x_w\|^2 - \|x_w\|^2\right| \leq \sum_{w=1}^{W} \varepsilon_w \|x_w\|^2 = \left(\sum_{w=1}^{W} \alpha_w(x)\varepsilon_w\right) \|x\|^2.$$

The weighted distortion is at most $\max_w \varepsilon_w$, since the weights sum to one. For independent blocks, the layer success events are independent, so their intersection has probability at least $\prod_w (1-\delta_w)$. □

The same block identities give the properties of the lifted map $S^\top S$ needed in the optimization analysis. None of the next three results requires independence between layers.

Lemma C.13 (Per-layer unbiasedness)**.** *If $\mathbb{E}[S_w^\top S_w] = I_{n_w}$ for every layer, then*

$$\mathbb{E}[S^\top S] = I_n, \qquad \mathbb{E}[S^\top S x] = x \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof. Take expectations in the block-diagonal identity for $S^\top S$. □

Lemma C.14 (Per-layer second moment)**.** *Suppose, for each layer, there is $\omega_w \geq 0$ such that*

$$\mathbb{E}\|S_w^\top S_w u\|^2 \leq \omega_w \|u\|^2 \qquad \textit{for every } u \in \mathbb{R}^{n_w}.$$

Then, for every $x \neq 0$,

$$\mathbb{E}\|S^\top S x\|^2 \leq \left(\sum_{w=1}^{W} \alpha_w(x)\omega_w\right) \|x\|^2 \leq \left(\max_w \omega_w\right) \|x\|^2.$$

Proof. Using the sum of squared block norms and linearity of expectation gives

$$\mathbb{E}\|S^\top S x\|^2 = \sum_{w=1}^{W} \mathbb{E}\|S_w^\top S_w x_w\|^2 \leq \sum_{w=1}^{W} \omega_w \|x_w\|^2.$$

Divide by $\|x\|^2$ and use $\sum_w \alpha_w(x) = 1$. □

Lemma C.15 (Per-layer high-probability norm bound)**.** *Suppose $\mathbb{P}[\|S_w\|^2 \leq c_w] \geq 1 - \delta_w$ for every layer. Then, with probability at least $1 - \sum_w \delta_w$,*

$$\|S\|^2 = \|S^\top S\| \leq \max_w c_w,$$

and, on the same event, for every $x \neq 0$,

$$\|Sx\|^2 \leq \left(\sum_{w=1}^{W} \alpha_w(x) c_w\right) \|x\|^2,$$
$$\|S^\top Sx\|^2 \leq \left(\sum_{w=1}^{W} \alpha_w(x) c_w^2\right) \|x\|^2.$$

Proof. A union bound ensures that all inequalities $\|S_w\|^2 \leq c_w$ hold on an event of the stated probability. On this event, block diagonality gives

$$\|S\|^2 = \|S^\top S\| = \max_w \|S_w\|^2 \leq \max_w c_w.$$

For any x, summing the layer bounds gives

$$\|Sx\|^2 \leq \sum_w c_w \|x_w\|^2, \qquad \|S^\top Sx\|^2 \leq \sum_w \|S_w\|^4 \|x_w\|^2 \leq \sum_w c_w^2 \|x_w\|^2.$$

Substituting $\|x_w\|^2 = \alpha_w(x)\|x\|^2$ proves the claim. □

Allocating the sketch dimensions. Theorem C.12 shows that a layer containing little of a vector's energy can tolerate a larger relative distortion. For Gaussian and Rademacher sketches, this allows a smaller sketch dimension, since k_w scales as ε_w^{-2}. The following allocation uses this weighted distortion budget for a fixed vector.

Corollary C.16 (Energy-aware allocation)**.** *Fix $x \neq 0$ and $\varepsilon, \delta \in (0,1)$. For every layer with $\alpha_w(x) > 0$, choose a Gaussian or Rademacher sketch with*

$$\varepsilon_w = \frac{\varepsilon\, \alpha_w(x)^{-1/3}}{\sum_{j=1}^{W} \alpha_j(x)^{2/3}}, \qquad \delta_w = \frac{\delta}{W}, \qquad k_w = \left\lceil C \frac{\log(2W/\delta)}{\varepsilon_w^2} \right\rceil,$$

where $C > 0$ is a sufficiently large universal constant. Assume that $\varepsilon_w < 1$ and $k_w \leq n_w$ for all such layers. Layers with $x_w = 0$ are omitted, with $k_w = 0$.

Then, for this fixed vector,

$$\mathbb{P}\left[\left|\|Sx\|^2 - \|x\|^2\right| \leq \varepsilon \|x\|^2\right] \geq 1 - \delta,$$

and the total number of transmitted coordinates satisfies

$$\sum_{w=1}^{W} k_w = \Theta\left(\frac{\log(2W/\delta)}{\varepsilon^2} \left(\sum_{w=1}^{W} \alpha_w(x)^{2/3}\right)^3\right).$$

Proof. The one-layer JL guarantees and a union bound give simultaneous success on all nonzero layers with probability at least $1 - \delta$. Zero layers contribute no error for the fixed vector x. The total relative distortion on this event is at most

$$\sum_{w:\alpha_w(x)>0} \alpha_w(x)\varepsilon_w = \varepsilon \frac{\sum_{w:\alpha_w(x)>0} \alpha_w(x)^{2/3}}{\sum_{j=1}^{W} \alpha_j(x)^{2/3}} = \varepsilon,$$

as in Theorem C.12.

To count coordinates, substitute the chosen distortions:

$$\sum_{w:\alpha_w(x)>0} \frac{1}{\varepsilon_w^2} = \frac{\left(\sum_{j=1}^{W} \alpha_j(x)^{2/3}\right)^2}{\varepsilon^2} \sum_{w:\alpha_w(x)>0} \alpha_w(x)^{2/3} = \frac{1}{\varepsilon^2} \left(\sum_{w=1}^{W} \alpha_w(x)^{2/3}\right)^3.$$

Since $0 < \varepsilon_w < 1$ and $\log(2W/\delta) \geq \log 2$, rounding each nonzero layer's dimension changes it by at most a universal constant factor. Multiplying the last identity by $\log(2W/\delta)$ therefore gives the stated total dimension. □

The allocation depends on the energy profile of x: its guarantee is for that fixed vector, with these dimensions chosen before the sketches are drawn. It applies only in the stated feasible regime $\varepsilon_w < 1$ and $k_w \leq n_w$.

C.6 KRONECKER SKETCHES: COMPRESSING ONE TENSOR MODE AT A TIME

Per-layer sketches act on separate blocks. A Kronecker sketch instead applies several smaller matrices to different modes of the same tensor. We use the notation from the main text: for $X \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and $S_j \in \mathbb{R}^{k_j \times n_j}$, define

$$\mathcal{S}(X) = X \times_1 S_1 \cdots \times_d S_d, \qquad S = S_d \otimes \cdots \otimes S_1.$$

Here $X \times_j S_j$ means multiplication along the jth mode. More explicitly, if $X_{(j)}$ is the matrix formed by arranging the mode-j fibers of X as columns, then

$$(X \times_j S_j)_{(j)} = S_j X_{(j)}, \qquad \operatorname{vec}(\mathcal{S}(X)) = S \operatorname{vec}(X).$$

Both unfolding and vectorization preserve the Frobenius norm. Thus a single mode multiplication can be analyzed using a matrix Frobenius-norm bound. We prove this composition step once, then apply it to dense sketches and to general JL factors.

Theorem C.17 (Combining independent Frobenius-norm guarantees)**.** *Let X be a fixed tensor and let $S_1, \ldots, S_d$ be independent random matrices. Fix $\varepsilon \in (0,1)$ and $\delta_1, \ldots, \delta_d \in [0,1]$. Suppose that, for each j and every fixed matrix V with n_j rows,*

$$\mathbb{P}\left[\left|\|S_j V\|_F^2 - \|V\|_F^2\right| > \frac{\varepsilon}{4d}\|V\|_F^2\right] \le \delta_j.$$

Then

$$\mathbb{P}\left[\left|\|\mathcal{S}(X)\|_F^2 - \|X\|_F^2\right| > \varepsilon\|X\|_F^2\right] \le \sum_{j=1}^{d} \delta_j.$$

Proof. Let $X^{(0)} = X$ and $X^{(j)} = X^{(j-1)} \times_j S_j$ denote the tensor after the first j multiplications. Conditional on $S_1, \ldots, S_{j-1}$, the tensor $X^{(j-1)}$ is fixed and S_j remains independent. The assumption applied to $(X^{(j-1)})_{(j)}$ therefore gives

$$\mathbb{P}\left[\left|\|X^{(j)}\|_F^2 - \|X^{(j-1)}\|_F^2\right| > \frac{\varepsilon}{4d}\|X^{(j-1)}\|_F^2 \;\middle|\; S_1, \ldots, S_{j-1}\right] \le \delta_j.$$

Taking expectations and a union bound, all d steps satisfy their bounds with probability at least $1 - \sum_j \delta_j$. On this event,

$$\left(1 - \frac{\varepsilon}{4d}\right)^d \|X\|_F^2 \le \|\mathcal{S}(X)\|_F^2 \le \left(1 + \frac{\varepsilon}{4d}\right)^d \|X\|_F^2.$$

For $0 < \varepsilon < 1$, Bernoulli's inequality and $1 + u \le e^u$ give

$$\left(1 - \frac{\varepsilon}{4d}\right)^d \ge 1 - \frac{\varepsilon}{4} \ge 1 - \varepsilon, \qquad \left(1 + \frac{\varepsilon}{4d}\right)^d \le e^{\varepsilon/4} \le 1 + \varepsilon.$$

This proves the claim. The proof also shows that it is enough to bound the conditional failure probability at each step; a simultaneous guarantee over all possible intermediate tensors is not required. □

Theorem C.18 (Gaussian and Rademacher Kronecker sketches)**.** *Let the factors $S_j \in \mathbb{R}^{k_j \times n_j}$ be independent Gaussian or Rademacher JL matrices. There is a universal constant $c > 0$ such that, for every fixed tensor X and every $\varepsilon \in (0,1)$,*

$$\mathbb{P}\left[\left|\|\mathcal{S}(X)\|_F^2 - \|X\|_F^2\right| > \varepsilon\|X\|_F^2\right] \le 2\sum_{j=1}^{d} e^{-c\varepsilon^2 k_j/d^2} \le 2d\, e^{-c\varepsilon^2 \min_j k_j/d^2}.$$

Consequently, failure probability at most δ is ensured by

$$\min_{1 \le j \le d} k_j \ge C\frac{d^2}{\varepsilon^2}\log\frac{2d}{\delta},$$

where $C > 0$ is a sufficiently large universal constant. The same bounds hold for $\|Sx\|^2$ with $x = \operatorname{vec}(X)$.

Proof. Lemma C.8, with distortion $\varepsilon/(4d)$, bounds the failure probability at step j by $2e^{-c_1 k_j \varepsilon^2/(16d^2)}$. Apply the conditioning argument of Theorem C.17 and set $c = c_1/16$. The dimension bound follows by requiring each of the d terms to be at most δ/d. Vectorization gives the final assertion. □

When only a fixed-vector JL guarantee is available. The preceding dense-sketch result uses the Frobenius-norm estimate of Lemma C.8, whose failure probability is independent of the matrix rank. For another construction, we can obtain a Frobenius-norm guarantee from fixed-vector bounds, at the cost of a rank factor.

Lemma C.19 (From fixed vectors to a fixed matrix)**.** *Fix $\varepsilon \in (0,1)$ and $\delta \in [0,1]$. Let $S \in \mathbb{R}^{k\times n}$ be a random matrix such that, for every fixed vector $u \in \mathbb{R}^n$,*

$$\mathbb{P}\left[\left|\|Su\|^2 - \|u\|^2\right| > \varepsilon\|u\|^2\right] \le \delta.$$

Then, for every fixed matrix V with n rows,

$$\mathbb{P}\left[\left|\|SV\|_F^2 - \|V\|_F^2\right| > \varepsilon\|V\|_F^2\right] \le \operatorname{rank}(V)\,\delta.$$

Proof. The assertion is immediate for $V = 0$. Otherwise, let u_i be its left singular vectors and $\sigma_i > 0$ its nonzero singular values. Then

$$\|SV\|_F^2 = \sum_i \sigma_i^2 \|Su_i\|^2, \qquad \|V\|_F^2 = \sum_i \sigma_i^2.$$

There are $\operatorname{rank}(V)$ unit vectors u_i. A union bound ensures that all their squared norms are preserved within ε with failure probability at most $\operatorname{rank}(V)\delta$. Multiplying their bounds by σ_i^2 and summing proves the claim. □

Theorem C.20 (Kronecker sketches from fixed-vector guarantees)**.** *Fix $\varepsilon \in (0,1)$ and $\delta_1, \dots, \delta_d \in [0,1]$. Let X be fixed and let $S_1, \dots, S_d$ be independent. Suppose that, for each j and every fixed $u \in \mathbb{R}^{n_j}$,*

$$\mathbb{P}\left[\left|\|S_j u\|^2 - \|u\|^2\right| > \frac{\varepsilon}{4d}\|u\|^2\right] \le \delta_j.$$

Then

$$\mathbb{P}\left[\left|\|\mathcal{S}(X)\|_F^2 - \|X\|_F^2\right| > \varepsilon\|X\|_F^2\right] \le \sum_{j=1}^{d} \operatorname{rank}(X_{(j)})\,\delta_j.$$

Proof. Use the intermediate tensors $X^{(j)}$ from Theorem C.17. The first $j-1$ mode multiplications only form linear combinations of columns in the mode-j unfolding. Hence

$$\operatorname{rank}\big((X^{(j-1)})_{(j)}\big) \le \operatorname{rank}(X_{(j)}).$$

Conditional on the first $j-1$ factors, Lemma C.19 therefore bounds the failure probability of step j by $\operatorname{rank}(X_{(j)})\delta_j$. The same union bound and multiplication of stepwise norm bounds complete the proof. □

The two composition theorems can be used at any desired distortion ε: evaluate each factor's available failure bound at $\varepsilon/(4d)$ and substitute it for δ_j. For Gaussian and Rademacher factors, the direct Frobenius bound gives the stronger, rank-independent result above.

The lifted operator. Finally, the algebra of Kronecker products transfers unbiasedness and operator-norm bounds from the factors to the full sketch.

Lemma C.21 (Unbiasedness and norm of Kronecker sketches)**.** *For independent factors $S_j \in \mathbb{R}^{k_j\times n_j}$, let $S = S_d \otimes \cdots \otimes S_1$. If $\mathbb{E}[S_j^\top S_j] = I_{n_j}$ for every j, then*

$$\mathbb{E}[S^\top S] = I_{n_1\cdots n_d}.$$

For every realization,

$$\|S\| = \prod_{j=1}^{d} \|S_j\|.$$

In particular, whenever $\|S_j\|^2 \le c_j$ for all j,

$$\|S\|^2 = \|S^\top S\| \le \prod_{j=1}^{d} c_j, \qquad \|S^\top S x\|^2 \le \left(\prod_{j=1}^{d} c_j\right)^2 \|x\|^2 \quad \text{for all } x.$$

Proof. The mixed-product identity gives

$$S^\top S = (S_d^\top S_d) \otimes \cdots \otimes (S_1^\top S_1).$$

Independence lets us take expectations factor by factor, proving unbiasedness. The singular values of a Kronecker product are products of the singular values of its factors, which proves the operator-norm identity. The remaining bounds follow from $\|S^\top S\| = \|S\|^2$ and $\|S^\top Sx\| \le \|S^\top S\|\|x\|$. $\square$

These results separate the ingredients needed later: norm and inner-product preservation control geometry in the compressed space, while moments and operator norms control the lifted update. A structured construction can be used in the high-probability convergence theorem once its parameters provide all three sketch events specified in Appendix D.

D BYZANTINE-ROBUST ANALYSIS

We prove convergence for fixed realized local and trusted objectives satisfying Assumptions 2.1, 2.9, and 2.10. The constants in these assumptions describe the optimization instance directly; no statistical model for the trusted dataset is imposed.

D.1 DESCENT AND CONVERGENCE

Let $\mathcal{F}_t$ denote the history before sampling S_t, including the fixed objectives, the current iterate, and all previous randomness and messages, and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. The gradients $g_0^t = \nabla f_0(x^t)$, $g^t = \nabla f(x^t)$, and $g_i^t = \nabla f_i(x^t)$ for $i \in \mathcal{G}$ are $\mathcal{F}_t$-measurable. The sketch S_t is independent of $\mathcal{F}_t$, while Byzantine messages may depend on the current sketch and honest messages. No worker has access to future sketch randomness. All probabilities and expectations below keep the objectives and the initial iterate fixed.

We first derive an estimate for the scalar product that appears in both convergence proofs.

Lemma D.1. *Let $x \in \mathbb{R}^n$ satisfy the reference-accuracy inequality in Assumption 2.10. Then, for every $\varepsilon \in [0, 1)$,*

$$\langle \nabla f_0(x), \nabla f(x) \rangle - \varepsilon \|\nabla f_0(x)\| \|\nabla f(x)\| \geq \frac{1 - \eta_1 - \varepsilon^2}{2} \|\nabla f(x)\|^2 - \frac{\eta_0}{2}. \tag{14}$$

Proof. Let $g_0 = \nabla f_0(x)$ and $g = \nabla f(x)$. By completing the square,

$$\begin{aligned} \langle g_0, g \rangle - \varepsilon \|g_0\| \|g\| &= \frac{1}{2} \left(\|g_0\| - \varepsilon \|g\| \right)^2 + \frac{1}{2} \left((1 - \varepsilon^2) \|g\|^2 - \|g_0 - g\|^2 \right) \\ &\geq \frac{1 - \eta_1 - \varepsilon^2}{2} \|g\|^2 - \frac{\eta_0}{2}, \end{aligned}$$

where the last inequality follows from Assumption 2.10. □

Further, we establish the descent lemma used for convergence both in expectation and with high probability.

Lemma D.2. *Let f_0 have an L-Lipschitz gradient. Then every iteration of Algorithm 1 satisfies $f_0(x^{t+1}) \leq f_0(x^t)$ and*

$$f_0(x^{t+1}) \leq f_0(x^t) - \frac{\gamma\beta}{M} \left\langle g_0^t, \sum_{i \in \mathcal{G}} u_i^t \right\rangle + \frac{L\gamma^2\beta}{2M} \sum_{i \in \mathcal{G}} \|u_i^t\|^2,$$

where $u_i^t = S_t^\top y_i^t$. For honest workers, $u_i^t = S_t^\top S_t g_i^t$.

Proof. The initial weights are non-negative and sum to one, and both branches of the weight update preserve these properties. Set

$$v^t = \sum_{i=1}^M \omega_i^t \mathbb{I}_{[\theta_i^t > 0]} u_i^t, \qquad q_i^t = [\theta_i^t]_0, \qquad Q^t = \sum_{i=1}^M q_i^t.$$

Applying convexity of the squared norm to the vectors $\mathbb{I}_{[\theta_i^t > 0]} u_i^t$ with weights ω_i^t gives

$$\|v^t\|^2 \leq \sum_{i=1}^M \omega_i^t \mathbb{I}_{[\theta_i^t > 0]} \|u_i^t\|^2.$$

Applying smoothness to the actual aggregated update, we obtain

$$\begin{aligned} f_0(x^{t+1}) &\leq f_0(x^t) - \gamma \langle g_0^t, v^t \rangle + \frac{L\gamma^2}{2} \|v^t\|^2 \leq f_0(x^t) - \gamma \sum_{i=1}^M \omega_i^t \mathbb{I}_{[\theta_i^t > 0]} \left(\langle g_0^t, u_i^t \rangle - \frac{L\gamma}{2} \|u_i^t\|^2 \right) \\ &= f_0(x^t) - \gamma \sum_{i=1}^M \omega_i^t q_i^t \leq f_0(x^t). \end{aligned}$$

If $Q^t > 0$, the weight update gives

$$\sum_{i=1}^M \omega_i^t q_i^t = (1 - \beta) \sum_{i=1}^M \omega_i^{t-1} q_i^t + \beta \frac{\sum_{i=1}^M (q_i^t)^2}{Q^t} \geq \beta \frac{\sum_{i=1}^M (q_i^t)^2}{Q^t} \geq \frac{\beta Q^t}{M} \geq \frac{\beta}{M} \sum_{i \in \mathcal{G}} \theta_i^t.$$

The first inequality drops a non-negative term. The second follows from $(Q^t)^2 \le M\sum_i (q_i^t)^2$, and the last uses $Q^t \ge \sum_{i\in\mathcal{G}} q_i^t \ge \sum_{i\in\mathcal{G}} \theta_i^t$. If $Q^t = 0$, then $v^t = 0$ and $\sum_{i\in\mathcal{G}} \theta_i^t \le 0$. Thus, in both cases,

$$f_0(x^{t+1}) \le f_0(x^t) - \frac{\gamma\beta}{M}\sum_{i\in\mathcal{G}} \theta_i^t.$$

Substituting $\theta_i^t = \langle g_0^t, u_i^t\rangle - \frac{L\gamma}{2}\|u_i^t\|^2$ proves the claimed bound. The argument uses smoothness of f_0 and convexity of the squared norm, without assuming convexity of f_0. □

We are now ready to present the convergence results. First, we derive convergence in expectation. To cover the different sketches mentioned above, we introduce two assumptions.

Assumption D.3 (Unbiasedness)**.** The lifted operator is conditionally unbiased if, for every $\mathcal{F}_t$-measurable vector $v \in \mathbb{R}^n$,

$$\mathbb{E}_t[S_t^\top S_t v] = v.$$

Assumption D.4 (Second moment boundedness in expectation)**.** There exists $\Omega \ge 1$ such that, for every $\mathcal{F}_t$-measurable vector $v \in \mathbb{R}^n$,

$$\mathbb{E}_t\|S_t^\top S_t v\|^2 \le \Omega\|v\|^2.$$

Theorem D.5. *Suppose that Assumptions 2.1, 2.9, 2.10, D.3, and D.4 hold. Let $T \ge 1$ be an integer, $\beta \in (0,1]$, $\Delta_0 = f_0(x^0) - f_0^{\inf}$, and choose $\gamma > 0$ such that*

$$D_{\mathrm{E}} \overset{def}{=} 1 - \eta_1 - L\gamma\Omega(1+\delta_2) > 0.$$

Then, after T iterations of Algorithm 1,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|^2 \le \frac{2M\Delta_0}{\gamma\beta G D_{\mathrm{E}} T} + \frac{\eta_0 + L\gamma\Omega\delta_1}{D_{\mathrm{E}}}. \tag{15}$$

In particular, if $0 < \gamma \le \frac{1-\eta_1}{2L\Omega(1+\delta_2)}$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|^2 \le \frac{4M\Delta_0}{\gamma\beta G(1-\eta_1)T} + \frac{2\eta_0}{1-\eta_1} + \frac{2L\gamma\Omega\delta_1}{1-\eta_1}. \tag{16}$$

Proof. Taking conditional expectation in Lemma D.2 and using $\frac{1}{G}\sum_{i\in\mathcal{G}} g_i^t = g^t$, we obtain

$$\begin{aligned}
\mathbb{E}_t f_0(x^{t+1}) &\le f_0(x^t) - \frac{\gamma\beta G}{M}\langle g_0^t, g^t\rangle + \frac{L\gamma^2\beta\Omega}{2M}\sum_{i\in\mathcal{G}}\|g_i^t\|^2 \\
&\le f_0(x^t) - \frac{\gamma\beta G}{2M}\left((1-\eta_1)\|g^t\|^2 - \eta_0\right) + \frac{L\gamma^2\beta\Omega G}{2M}\left((1+\delta_2)\|g^t\|^2 + \delta_1\right) \\
&= f_0(x^t) - \frac{\gamma\beta G}{2M}\left(D_{\mathrm{E}}\|g^t\|^2 - \eta_0 - L\gamma\Omega\delta_1\right).
\end{aligned}$$

Here, the second inequality uses Lemma D.1 with $\varepsilon = 0$ and equation 7: Assumption 2.10 controls alignment, while Assumption 2.9 bounds the honest squared norms. Taking full expectation, summing over $t = 0, \dots, T-1$, and using $f_0(x^T) \ge f_0^{\inf}$ gives equation 15. Under the displayed stepsize restriction, $D_{\mathrm{E}} \ge (1-\eta_1)/2$, which proves the last bound. □

For fresh Gaussian and Rademacher matrices, Lemmas 2.6 and 2.7 give $\Omega = (n+k+1)/k$ and $\Omega = (n+k-1)/k$, respectively. The expectation is over the subsequent algorithmic randomness for fixed objectives satisfying the stated assumptions.

We now derive convergence with high probability. The proof needs three events for each fresh sketch: preservation of one inner product, control of the total compressed honest energy, and an operator-norm bound.

Theorem D.6 (Theorem 2.11 from the main part)**.** *Under the assumptions and notation of Theorem 2.11, except for its stepsize restriction, define*

$$c_\gamma \overset{def}{=} L\gamma\rho(1+\varepsilon), \qquad D_\gamma \overset{def}{=} d - (1+\delta_2)c_\gamma.$$

More generally, for every $\gamma > 0$ with $D_\gamma > 0$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x^t)\|^2 \le \frac{2M\Delta_0}{\gamma\beta G D_\gamma T} + \frac{\eta_0 + c_\gamma\delta_1}{D_\gamma} \tag{17}$$

with probability at least $1-\delta$*. The stepsize in Theorem 2.11 gives equation 11.*

Proof. Set $q = \delta/(4T)$ and let $V_t \in \mathbb{R}^{n\times G}$ be the matrix whose columns are the honest gradients g_i^t. Conditional on $\mathcal{F}_t$, both V_t and the pair (g_0^t, g^t) are fixed. Lemma C.8 gives

$$\mathbb{P}\left[\|S_tV_t\|_F^2 > (1+\varepsilon)\|V_t\|_F^2 \mid \mathcal{F}_t\right] \le 2e^{-c_1k\varepsilon^2} \le q$$

when $k \ge C\varepsilon^{-2}\log(8T/\delta)$ for a sufficiently large universal constant C. The same lemma for a single column gives the fixed-vector JL bound with failure probability at most q. Applying Lemma C.1 then gives

$$\left|\left\langle S_tg_0^t, S_tg^t\right\rangle - \left\langle g_0^t, g^t\right\rangle\right| \le \varepsilon\|g_0^t\|\|g^t\|$$

with conditional failure probability at most $2q$. Finally, Lemma 2.8 gives $\|S_t\|^2 \le \rho$ with conditional failure probability at most q. Thus, with conditional probability at least $1-4q$, all three events hold.

On their intersection, Lemma D.1 implies

$$\left\langle g_0^t, \frac{1}{G}\sum_{i\in\mathcal{G}} u_i^t \right\rangle = \left\langle S_tg_0^t, S_tg^t\right\rangle \ge \left\langle g_0^t, g^t\right\rangle - \varepsilon\|g_0^t\|\|g^t\| \ge \frac{d}{2}\|g^t\|^2 - \frac{\eta_0}{2}.$$

Moreover, using the operator norm and the Frobenius norm in this order,

$$\begin{aligned}\sum_{i\in\mathcal{G}}\|u_i^t\|^2 &= \|S_t^\top S_tV_t\|_F^2 \le \|S_t\|^2\|S_tV_t\|_F^2 \le \rho(1+\varepsilon)\|V_t\|_F^2 \\ &\le \rho(1+\varepsilon)G\left((1+\delta_2)\|g^t\|^2 + \delta_1\right).\end{aligned}$$

Substituting these two estimates into Lemma D.2 yields

$$\begin{aligned}f_0(x^{t+1}) &\le f_0(x^t) - \frac{\gamma\beta G}{2M}\left(d\|g^t\|^2 - \eta_0\right) + \frac{L\gamma^2\beta\rho(1+\varepsilon)G}{2M}\left((1+\delta_2)\|g^t\|^2 + \delta_1\right) \\ &= f_0(x^t) - \frac{\gamma\beta G}{2M}\left(D_\gamma\|g^t\|^2 - \eta_0 - c_\gamma\delta_1\right).\end{aligned}$$

Let $\mathcal{E}_t$ be the intersection of the three sketch events at iteration t. Although the iterates are adaptive, their conditional failure probabilities are bounded for every history. Therefore,

$$\mathbb{P}\left[\bigcup_{t=0}^{T-1}\mathcal{E}_t^c\right] \le \sum_{t=0}^{T-1}\mathbb{E}\left[\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_t)\right] \le 4Tq = \delta.$$

On $\bigcap_{t=0}^{T-1}\mathcal{E}_t$, summing the descent inequality and using $f_0(x^T) \ge f_0^{\inf}$ gives

$$\frac{\gamma\beta GD_\gamma}{2M}\sum_{t=0}^{T-1}\|g^t\|^2 \le \Delta_0 + \frac{\gamma\beta GT}{2M}\left(\eta_0 + c_\gamma\delta_1\right).$$

Dividing by $\gamma\beta GD_\gamma T/(2M)$ proves equation 17. Finally, the stepsize in Theorem 2.11 gives $D_\gamma \ge d/2$, which proves equation 11. □

The probability estimates involve only honest gradients and the trusted reference, all fixed before the fresh sketch is sampled. No geometric preservation event is imposed on Byzantine messages. The same proof applies to another sketch construction if it provides the three conditional events above with the stated failure probabilities and a suitable value of ρ; the required dimension must then be established for that construction.