LASER: Linear Compression in Wireless Distributed Optimization

Ashok Vardhan Makkuva *, **Marco Bondaschi** *, **Thijs Vogels, Martin Jaggi**
School of Computer and Communication Sciences
EPFL
Lausanne, Switzerland
{ashok.makkuva, marco.bondaschi, thijs.vogels, martin.jaggi}@epfl.ch

Hyeji Kim
Department of Electrical and Computer Engineering
UT Austin
Austin, TX, USA
hyeji.kim@austin.utexas.edu

Michael Gastpar
School of Computer and Communication Sciences
EPFL
Lausanne, Switzerland
michael.gastpar@epfl.ch

Abstract

Data-parallel SGD is the de facto algorithm for distributed optimization, especially for large scale machine learning. Despite its merits, communication bottleneck is one of its persistent issues. Most compression schemes to alleviate this either assume noiseless communication links, or fail to achieve good performance on practical tasks. In this paper, we close this gap and introduce LASER: **L**ine**A**r Compre**S**sion in Wir**E**less Dist**R**ibuted Optimization. LASER capitalizes on the inherent low-rank structure of gradients and transmits them efficiently over the noisy channels. Whilst enjoying theoretical guarantees similar to those of the classical SGD, LASER shows consistent gains over baselines on a variety of practical benchmarks. In particular, it outperforms the state-of-the-art compression schemes on challenging computer vision and GPT language modeling tasks. On the latter, we obtain 50-64% improvement in perplexity over our baselines for noisy channels.

1 Introduction

Distributed optimization is one of the most widely used frameworks for training large scale deep learning models [1, 2, 3]. In particular, data-parallel SGD is the workhorse algorithm for this task. Underpinning this approach is the *communication* of large gradient vectors between the workers and the central server which performs their *aggregation*. While these methods harness the inherent parallelism to reduce the overall training time, their communication cost is a major bottleneck that limits scalability to large models. Design of communication-efficient distributed algorithms is thus a must for reaping the full benefits of distributed optimization [4].

*Equal contribution.

International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023 (FL@FM-NeurIPS'23). This workshop does not have official proceedings and this paper is non-archival.

Existing approaches to reduce the communication cost can be broadly classified into two themes: (i) compressing the gradients before transmission; or (ii) utilizing the communication link for native 'over-the-air' aggregation (averaging) across workers. Along (i), a number of gradient compression schemes have been designed such as quantization [5, 6], sparsification [7, 8], hybrid methods [9, 10], and low-rank compression [11, 12]. These methods show gains over the full-precision SGD in various settings ([4] is a detailed survey). Notwithstanding the merits, their key shortcoming is that they assume a *noiseless* communication link between the clients and the server. In settings such as federated learning with differential privacy or wireless communication, these links are noisy. Making them noiseless requires error-correcting codes which exacerbates the latency, as the server needs to wait till it receives the gradient from each worker before aggregating [13].

Under theme (ii), communication cost is reduced by harnessing the physical layer aspects of (noisy) communication. In particular, the superposition nature of wireless channels is exploited to perform over-the-air averaging of gradients across workers, which reduces the latency, see e.g. [14] and the references therein. Notable works include A-DSGD [15], analog-gradient-aggregation [13, 16], channel aware quantization [17], etc. However, to the best of our knowledge, the majority of these approaches are restricted to synthetic datasets and shallow neural networks (often single layer) and do not scale well to the practical neural network models (which we verify in Sec. 4). This leads to a natural question:

Can we design efficient and practical gradient compression schemes for noisy communication channels?

In this work, we precisely address this and propose LASER, a principled gradient compression scheme for distributed training over wireless noisy channels. Specifically, we make the following contributions:

- Capitalizing on the inherent low-rank structure of the gradients, LASER efficiently computes these low-rank factors and transmits them reliably over the noisy channel while allowing the gradients to be averaged in transit (Sec. 3).
- We show that LASER enjoys similar convergence rate as that of the classical SGD for both quasi-convex and non-convex functions, except for a small additive constant depending on the channel degradation (Theorem 1).
- We empirically demonstrate the superiority of LASER over the baselines on the challenging tasks of (i) language modeling with GPT-2 $\to$ WIKITEXT-103 and (ii) image classification with RESNET18 $\to$ (CIFAR10, CIFAR100) and 1-LAYER NN $\to$ MNIST. With high gradient compression (165×), LASER achieves 50-64% perplexity improvement in the low and moderate power regimes on WIKITEXT-103. To the best of our knowledge, LASER is the first to exhibit such gains for GPT language modeling (Sec. 4).

Notation. Euclidean vectors and matrices are denoted by bold letters $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}$, etc. $\|\cdot\|$ denotes the Frobenius norm for matrices and the ℓ_2-norm for Euclidean vectors. $\mathcal{O}(\cdot)$ is an upper bound subsuming universal constants whereas $\widetilde{\mathcal{O}}(\cdot)$ hides any logarithmic problem-variable dependencies.

2 Background

Distributed optimization. Consider the (synchronous) data-parallel distributed setting where we minimize an objective $f : \mathbb{R}^d \to \mathbb{R}$ defined as the empirical loss on a global dataset $\mathcal{D} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^N$:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\theta}), \quad f(\boldsymbol{\theta}) \triangleq \frac{1}{N} \sum_{j=1}^{N} \ell(\boldsymbol{x}_j, y_j; \boldsymbol{\theta}),$$

where $\ell(\cdot)$ evaluates the loss for each data sample $(\boldsymbol{x}_j, y_j)$ on model $\boldsymbol{\theta}$. In this setup, there are k (data-homogeneous) training clients, where the i^{th} client has access to a stochastic gradient oracle $\boldsymbol{g}_i$, e.g. mini-batch gradient on a set of samples randomly chosen from $\mathcal{D}$, such that $\mathbb{E}[\boldsymbol{g}_i|\boldsymbol{\theta}] = \nabla f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$. In distributed SGD [18, 1], the server aggregates all $\boldsymbol{g}_i$s and performs the following updates:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \cdot \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{g}_i^{(t)}, \quad \mathbb{E}[\boldsymbol{g}_i^{(t)}|\boldsymbol{\theta}_t] = \nabla f(\boldsymbol{\theta}_t), \quad t \geq 0, \tag{SGD}$$

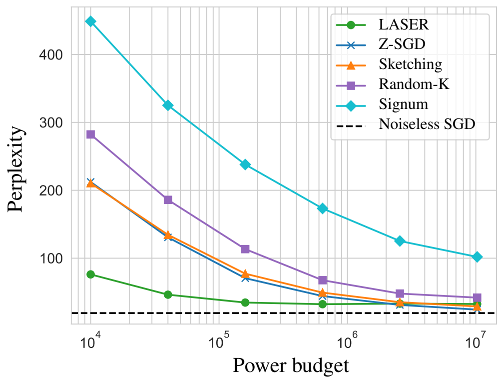

Figure 1: Final test perplexity after 20k iterations *(lower is better)* vs. power budget for GPT-2 language modeling on WIKITEXT-103. LASER consistently requires orders-of-magnitude less power than other methods for the same perplexity.

Table 1: Power required *(lower is better)* to reach the target perplexity on WIKITEXT-103. Z-SGD sends the uncompressed gradients directly, while LASER sends a rank-4 approximation. LASER requires 16× less power than Z-SGD to achieve the target perplexity over a wide interval. In the very-high-power regime with perplexity close to that of the noiseless SGD, we see no power gains.

Target	**Power required**		**Reduction**
	Z-SGD	LASER	
80	160 K	10 K	16×
50	640 K	40 K	16×
40	2560 K	160 K	16×
35	2560 K	160 K	16×

where $\{\gamma_t\}_{t\geq 0}$ is a stepsize schedule. Implicit here is the assumption that the communication link between the clients and the server is noiseless, which we expound upon next.

Communication model. For the communication uplink from the clients to the server, we consider the standard wireless channel for over-the-air distributed learning [19, 13, 16, 17, 20]: the *additive slow-fading channel*, e.g., the classical multiple-access-channel [21]. The defining property of this family is the superposition of incoming wireless signals (enabling over-the-air computation) possibly corrupted together with an independent channel noise [14]. Specifically, we denote the channel as a (random) mapping $\mathcal{Z}_P(\cdot)$ that transforms the set of (time-varying) messages transmitted by the clients $\{\boldsymbol{x}_i\}_{i\in[k]} \subset \mathbb{R}^d$ to its noisy version $\boldsymbol{y} \in \mathbb{R}^d$ received by the server:

$$\boldsymbol{y} = \mathcal{Z}_P(\{\boldsymbol{x}_i\}) \triangleq \sum_{i=1}^{k} \boldsymbol{x}_i + \boldsymbol{Z}, \quad \|\boldsymbol{x}_i\|^2 \leq P_t,\ \frac{1}{T}\sum_{t=0}^{T-1} P_t \leq P, \tag{1}$$

where the noise $\boldsymbol{Z} \in \mathbb{R}^d$ is independent of the channel inputs and has zero mean and unit variance per dimension, i.e. $\mathbb{E}\|\boldsymbol{Z}\|^2 = d$. The power constraint on each client $\|\boldsymbol{x}_i\|^2 \leq P_t$ at time t serves as a communication cost (and budget), while the power policy $\{P_t\}$ allots the total budget P over T epochs as per the average power constraint [22, 15]. A key metric that captures the channel degradation quality is the signal-to-noise ratio per coordinate (SNR), defined as the ratio between the average signal energy (P) and that of the noise (d), i.e. $\mathrm{SNR} \triangleq P/d$. The larger it is the better the signal fidelity. The power budget P encourages the compression of signals: if each client can transmit the same information $\boldsymbol{x}_i$ via fewer entries (smaller d), they can utilize more power per entry (higher SNR) and hence a more faithful signal. For the downlink communication from the server to the clients (broadcast channel), we assume that it is noiseless and thus the clients exactly receive what the server transmits [23, 24, 25]. In the rest of the paper by channel we mean the uplink channel. The channel model in Eq. (1) readily generalizes to the fast fading setup as discussed in Sec. 4.

Gradient transmission over the channel. In the distributed optimization setting the goal is to communicate the (time-varying) local gradients $\boldsymbol{g}_i \in \mathbb{R}^d$ to the central server over the noisy channel in Eq. (1). Here we set the messages $\boldsymbol{x}_i$ as linear scaling of gradients (as we want to estimate the gradient average), i.e. $\boldsymbol{x}_i = a_i\, \boldsymbol{g}_i$ with the scalars $a_i \in \mathbb{R}$ enforcing the power constraints:

$$\boldsymbol{y} = \sum_{i=1}^{k} a_i\, \boldsymbol{g}_i + \boldsymbol{Z}, \quad \|a_i\, \boldsymbol{g}_i\|^2 \leq P_t. \tag{2}$$

Now the received signal is a weighted sum of the gradients corrupted by noise, whereas we need the sum of the gradients $\sum_i \boldsymbol{g}_i$ (upto zero mean additive noise) for the model training. Towards this goal, a common mild technical assumption is that the gradient norms $\{\|\boldsymbol{g}_i\|\}$ are known at the receiver at each communication round [17, 13] (can be relaxed in practice, Sec. 4). The optimal scalars are then given by $a_i = \sqrt{P_t}/(\max_j \|\boldsymbol{g}_j\|), \forall i \in [K]$, which are uniform across all the clients (§ F.1). Now substituting this a_i in Eq. (2) and rearranging, the effective channel can be written as

$$\boldsymbol{y} = \widetilde{\mathcal{Z}}_P(\{\boldsymbol{g}_i\}) \triangleq \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{g}_i + \frac{\max_i \|\boldsymbol{g}_i\|}{k\sqrt{P_t}}\, \boldsymbol{Z}. \qquad \text{(noisy channel)}$$

Or equivalently, we can assume this as the actual channel model where the server receives the gradient average corrupted by a zero mean noise proportional to the gradients. Note that the noise magnitude decays in time as gradients converge to zero. We denote $\widetilde{\mathcal{Z}}_P(\cdot)$ as simply $\mathcal{Z}_P(\cdot)$ henceforth as these two mappings are equivalent.

Z-SGD. Recall that the SGD aggregates the uncompressed gradients directly. In the presence of the noisy channel, it naturally modifies to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t\, \mathcal{Z}_P(\{\boldsymbol{g}_i^{(t)}\}). \qquad \text{(Z-SGD)}$$

Thus Z-SGD is a canonical baseline to compare against. It has two sources of stochasticity: one stemming for the stochastic gradients and the other from the channel noise. While the gradient in the Z-SGD update still has the same conditional mean as the noiseless case (zero mean Gaussian in noisy channel), it has higher variance due to the Gaussian term. When $P = \infty$, Z-SGD reduces to SGD.

3 LASER: Novel linear compression cum transmission scheme

In this section we describe our main contribution, LASER, a novel method to compress gradients and transmit them efficiently over noisy channels. The central idea underpinning our approach is that, given the channel power constraint in Eq. (1), we can get a more faithful gradient signal at the receiver by transmitting its 'appropriate' compressed version (fewer entries sent and hence more power per entry) as opposed to sending the full-gradient naively as in Z-SGD. This raises a natural question: *what's a good compression scheme that facilitates this?* To address this, we posit that we can capitalize on the inherent low-rank structure of the gradient matrices [26, 27, 28] for efficient gradient compression and transmission. Indeed, as illustrated in Theorem 1, we can get a variance reduction of the order of the smaller dimension when the gradient matrices are approximately low-rank. Algorithm 1 below details LASER. While LASER works with any power policy $\{P_t\}$ in noisy channel, it suffices to consider the constant law $P_t = P$ as justified in Sec. 4.2

3.1 Algorithm

Algorithm 1 LASER: Linear compression in wireless distributed optimization

input: initial model parameters $\boldsymbol{\theta} \in \mathbb{R}^{m \times n}$, learning rate γ, compression rank r, power budget P
output: trained parameters $\boldsymbol{\theta}$
at each worker $i = 1, \dots, k$ **do**
  initialize memory $\boldsymbol{e}_i \leftarrow \boldsymbol{0} \in \mathbb{R}^{m \times n}$
  for each iterate $t = 0, \dots$ **do**
    Compute a stochastic gradient $\boldsymbol{g}_i \in \mathbb{R}^{m \times n}$
    $\boldsymbol{M}_i \leftarrow \boldsymbol{e}_i + \gamma \boldsymbol{g}_i$ ▷ Updated gradient via error feedback
    $\boldsymbol{P}_i, \boldsymbol{Q}_i \leftarrow \mathcal{C}_r(\boldsymbol{M}_i)$ ▷ Rank-r compression [12]
    $\boldsymbol{e}_i \leftarrow \boldsymbol{M}_i - \text{DECOMPRESS}(\mathcal{C}_r(\boldsymbol{M}_i))$ ▷ Memorize local errors
    $\boldsymbol{\alpha}, \boldsymbol{\beta} \leftarrow \text{POWERALLOC}(\{\mathcal{C}_r(\boldsymbol{M}_j), \boldsymbol{M}_j\})$ ▷ Power allocation
    $\boldsymbol{Y}_{\boldsymbol{p}},\ \boldsymbol{Y}_{\boldsymbol{q}} \leftarrow \mathcal{Z}_{\boldsymbol{\alpha}}(\{\boldsymbol{P}_j\}),\ \mathcal{Z}_{\boldsymbol{\beta}}(\{\boldsymbol{Q}_j\})$ ▷ Channel transmission
    $\boldsymbol{g} \leftarrow \text{DECOMPRESS}(\boldsymbol{Y}_{\boldsymbol{p}}, \boldsymbol{Y}_{\boldsymbol{q}})$ ▷ Reconstruct the gradient in $\mathbb{R}^{m \times n}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \boldsymbol{g}$
  end for
end at

For distributed training of neural network models, we apply Algorithm 1 to each layer independently, only for the weight matrices and the convolutional filters, and transmit the bias vectors uncompressed.

3.2 Theoretical results

We now provide theoretical justification for LASER for learning parameters in $\mathbb{R}^{m\times n}$ with $m \le n$ (without loss of generality). While our algorithm works for any number of clients, for the theory we consider $k = 1$ to illustrate the primary gains with our approach. Our results readily extend to the multiple clients setting following [29]. Specifically, Theorem 1 below highlights that the asymptotic convergence rate of LASER is *almost the same as that of the classical* SGD, except for a small additive constant $\lambda_{\rm LASER}$ which is $\mathcal{O}(m)$ times smaller than that of Z-SGD. Our results hold for both quasi-convex and arbitrary non-convex functions. We start with the preliminaries.

Definition 1 (**Channel influence factor**). *For any compression cum transmission algorithm* ALG*, let $\boldsymbol{y}_{\rm ALG}(\boldsymbol{g})$ be the reconstructed gradient at the server after transmitting $\boldsymbol{g}$ over the noisy channel. Then the channel influence factor $\lambda_{\rm ALG}$ is defined as*

$$\lambda_{\rm ALG} \triangleq \frac{\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{y}_{\rm ALG}(\boldsymbol{g}) - \boldsymbol{g}\|^2}{\|\boldsymbol{g}\|^2}. \tag{3}$$

The influence factor gauges the effect of the channel on the variance of the final gradient $\boldsymbol{y}_{\rm ALG}$: if the original stochastic gradient $\boldsymbol{g}$ has variance σ^2 with respect to the actual gradient ∇f, then $\boldsymbol{y}_{\rm ALG}$ has $(1+\lambda_{\rm ALG})\sigma^2$. Note that this variance directly affects the convergence speed of the SGD and hence the smaller $\lambda_{\rm ALG}$ is, the better the compression scheme is. In view of this, the following fact (§ C.2) illustrates the crucial gains of LASER compared to Z-SGD, which are roughly of order $\mathcal{O}(m)$ (§ F.3):

$$\lambda_{\rm LASER} \le \frac{4}{(m/r){\rm SNR}}\left(1 + \frac{1}{(n/r){\rm SNR}}\right) \ll \frac{1}{\rm SNR} = \lambda_{\text{Z-SGD}}. \tag{4}$$

We utilize the standard assumptions for SGD convergence following the framework in [1] and [30] and defer them to § B. We are now ready to state our main result.

Theorem 1 (**LASER convergence**). *Let $\{\boldsymbol{\theta}_t\}_{t\ge 0}$ be the* LASER *iterates (Alg. 1) with constant stepsize schedule $\{\gamma_t = \gamma\}_{t\ge 0}$ and suppose Assumptions 2-5 hold. Denote $\boldsymbol{\theta}_\star \triangleq \operatorname{argmin}_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$, $f_\star \triangleq f(\boldsymbol{\theta}_\star)$, and $\tau \triangleq 10L\left(\frac{2}{\delta_r} + M\right)$. Then for $k = 1$,*

(i) *if f is μ-quasi convex for $\mu > 0$, there exists a stepsize $\gamma \le \frac{1}{\tau(1+\lambda_{\rm LASER})}$ such that*

$$\mathbb{E}f(\boldsymbol{\theta}_{\rm out}) - f_\star = \widetilde{\mathcal{O}}\left(\tau(1+\lambda_{\rm LASER})\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|^2 \exp\left(\frac{-\mu T}{\tau(1+\lambda_{\rm LASER})}\right) + \frac{\sigma^2(1+\lambda_{\rm LASER})}{\mu T}\right),$$

where $\boldsymbol{\theta}_{\rm out}$ is chosen from $\{\boldsymbol{\theta}\}_{t=0}^{T-1}$ such that $\boldsymbol{\theta}_{\rm out} = \boldsymbol{\theta}_t$ with probability $(1-\mu\gamma/2)^{-t}$.

(ii) *if f is μ-quasi convex for $\mu = 0$, there exists a stepsize $\gamma \le \frac{1}{\tau(1+\lambda_{\rm LASER})}$ such that*

$$\mathbb{E}f(\boldsymbol{\theta}_{\rm out}) - f_\star = \mathcal{O}\left(\frac{\tau\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|^2(1+\lambda_{\rm LASER})}{T} + \sigma\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|\sqrt{\frac{1+\lambda_{\rm LASER}}{T}}\right),$$

where $\boldsymbol{\theta}_{\rm out}$ is chosen uniformly at random from $\{\boldsymbol{\theta}\}_{t=0}^{T-1}$.

(iii) *if f is an arbitrary non-convex function, there exists a stepsize $\gamma \le \frac{1}{\tau(1+\lambda_{\rm LASER})}$ such that*

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{\rm out})\|^2 = \mathcal{O}\left(\frac{\tau\|f(\boldsymbol{\theta}_0) - f_\star\|^2(1+\lambda_{\rm LASER})}{T} + \sigma\sqrt{\frac{L(f(\boldsymbol{\theta}) - f_\star)(1+\lambda_{\rm LASER})}{T}}\right),$$

where $\boldsymbol{\theta}_{\rm out}$ is chosen uniformly at random from $\{\boldsymbol{\theta}\}_{t=0}^{T-1}$.

(iv) Z-SGD *obeys the convergence bounds (i)-(iii) with $\delta_r = 1$ and $\lambda_{\rm LASER}$ replaced by $\lambda_{\text{Z-SGD}}$.*

LASER vs. Z-SGD. Thus the asymptotic rate of LASER is dictated by the timescale $(1 + \lambda_{\rm LASER})/T$, very close to the $1/T$ rate for the classical SGD. In contrast, Z-SGD has the factor $(1+\lambda_{\text{Z-SGD}})/T$ with $\lambda_{\text{Z-SGD}} = \mathcal{O}(m)\,\lambda_{\rm LASER}$.

Multiple clients. As all the workers in LASER (Alg. 1) apply the same linear operations for gradient compression (via PowerSGD), Theorem 1 can be extended to (homogenous) multiple workers by shrinking the constants σ^2, SNR, $\lambda_{\rm LASER}$, and $\lambda_{\text{Z-SGD}}$ by a factor of k, following [29].

Table 2: Benchmarks for evaluating LASER. Baseline refers to the noiseless SGD.

Task	Model	Params	Dataset	Metric	Baseline
Language modeling	GPT-2	123.6 M	WIKITEXT-103	Perplexity	19.2
Image classification	RESNET18	11.2 M 11.2 M	CIFAR10 CIFAR100	Top-1 accuracy	93.0% 73.1%
	1-LAYER NN	7850	MNIST		92.3%

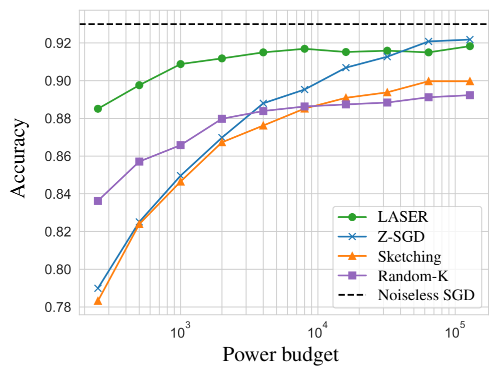

Figure 2: Test accuracy (*higher the better*) for a given power budget on CIFAR-10 for different algorithms.

Table 3: Communication cost (*lower the better*) for GPT language modeling on WIKITEXT-103. LASER transmits the lowest volume of data during training.

Algorithm	Data sent per iteration	
Z-SGD	496 MB	(1×)
SIGNUM	15 MB	(33×)
RANDOM-K	99 MB	(5×)
SKETCHING	99 MB	(5×)
A-DSGD	n/a	n/a
LASER	**3 MB**	(165×)

4 Experimental results

We empirically demonstrate the superiority of LASER over state-of-the-art baselines on a variety of benchmarks, summarized in Table 2. We defer further details to § G.

4.1 Results on language modeling and image classification

For GPT language modeling, Fig. 1 in Sec. 1 highlights that LASER outperforms the baselines over a wide range of power levels. To the best of our knowledge, this is the first result of its kind to demonstrate gains for GPT training over noisy channels. Specifically, we obtain 64% improvement in perplexity over Z-SGD (76 vs. 212) in the low power regime ($P = 10\,\mathrm{K}$) and 50% (35 vs. 71) for the moderate one ($P = 160\,\mathrm{K}$). This demonstrates the efficacy of LASER especially in the limited power environment. Indeed, Table 1 illustrates that for a fixed target perplexity, LASER requires 16× less power than the second best, Z-SGD. In the very high power regime, we observe no clear gains (as expected) compared to transmitting the uncompressed gradients directly via the Z-SGD.

We observe a similar trend for CIFAR10 classification, as Fig. 2 demonstrate the superiority of LASER over other compression schemes; SIGNUM is considerably worse than others and hence omitted. As for power reduction, we observe similar gains to Table 1 (see § G).

4.2 Power control: static vs. dynamic policies

The formulation in noisy channel allows for any power control law P_t as long as it satisfies the average power constraint: $\sum_t (P_t/T) \leq P$. This begs a natural question: *what's the best power scheme for LASER?* To answer this, for CIFAR10 classification, under a fixed budget P we consider different power policies with both increasing and decreasing power across epochs: the constant, piecewise constant and linear schemes. Fig. 3 illustrates the results for the decreasing power laws, while Fig. 7 their increasing counterparts. These results highlight that the *constant* power policy achieves the *best* per-

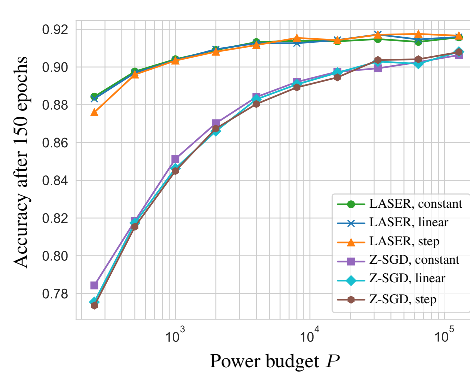

Figure 3: Accuracy vs. budget P for various laws. Constant is the best for both LASER and Z-SGD.

formance for both LASER and Z-SGD, compared to the time-varying ones. Further LASER attains significant accuracy gains over Z-SGD for all the power control laws. Interestingly LASER performs the *same* with all the power schemes. We posit this behavior to the fact that the noisy channel already contains a time-varying noise due to the term $\frac{\max_i |||\boldsymbol{g}_i|||}{\sqrt{P_t}}$. Since the gradients decay over time, this inherently allows for an implicit power/SNR-control law even with a constant P_t, thus enabling the constant power scheme to fare as good as the others. Hence, without loss of generality, we consider the static power schedule for our theory and experiments. We refer to § H.2 for a detailed discussion.

5 Conclusion

We propose a principled gradient compression scheme, LASER, for wireless distributed optimization over additive noise channels. LASER attains significant gains over its baselines on a variety of metrics such as accuracy/perplexity, complexity and communication cost. It is an interesting avenue of future research to extend LASER to channels with downlink noise and fast fading without CSI.

References

[1] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[2] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

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

[4] Hang Xu, Chen-Yu Ho, Ahmed M Abdelmoniem, Aritra Dutta, El Houcine Bergou, Konstantinos Karatsenidis, Marco Canini, and Panos Kalnis. Compressed communication for distributed deep learning: Survey and quantitative evaluation. Technical report, 2020.

[5] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*, 2018.

[6] Shay Vargaftik, Ran Ben Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben Itzhak, and Michael Mitzenmacher. Eden: Communication-efficient and robust distributed mean estimation for federated learning. In *International Conference on Machine Learning*, pages 21984–22014. PMLR, 2022.

[7] Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

[8] Berivan Isik, Francesco Pase, Deniz Gunduz, Tsachy Weissman, and Michele Zorzi. Sparse random networks for communication-efficient federated learning. *arXiv preprint arXiv:2209.15328*, 2022.

[9] Jiawei Jiang, Fangcheng Fu, Tong Yang, and Bin Cui. Sketchml: Accelerating distributed machine learning with data sketches. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1269–1284, 2018.

[10] Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in Neural Information Processing Systems*, 31, 2018.

[12] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Huayan Guo, An Liu, and Vincent KN Lau. Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis. *IEEE Internet of Things Journal*, 8(1):197–210, 2020.

[14] Yuanming Shi, Kai Yang, Tao Jiang, Jun Zhang, and Khaled B Letaief. Communication-efficient edge AI: Algorithms and systems. *IEEE Communications Surveys & Tutorials*, 22(4):2167–2191, 2020.

[15] Mohammad Mohammadi Amiri and Deniz Gündüz. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. *IEEE Transactions on Signal Processing*, 68:2155–2169, 2020.

[16] Guangxu Zhu, Yong Wang, and Kaibin Huang. Broadband analog aggregation for low-latency federated edge learning. *IEEE Transactions on Wireless Communications*, 19(1):491–506, 2019.

[17] Wei-Ting Chang and Ravi Tandon. Mac aware quantization for distributed gradient descent. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, 2020.

[18] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[19] Mohammad Mohammadi Amiri and Deniz Gündüz. Federated learning over wireless fading channels. *IEEE Transactions on Wireless Communications*, 19(5):3546–3557, 2020.

[20] Xizixiang Wei and Cong Shen. Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1253–1268, 2022.

[21] Bobak Nazer and Michael Gastpar. Computation over multiple-access channels. *IEEE Transactions on information theory*, 53(10):3498–3516, 2007.

[22] Xizixiang Wei and Cong Shen. Federated learning over noisy channels: Convergence analysis and design examples. *IEEE Transactions on Cognitive Communications and Networking*, 8(2):1253–1268, 2022.

[23] H Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *https://ai.googleblog.com/2017/04/federated-learning-collaborative.html*, 2017.

[24] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[25] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[26] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *The Journal of Machine Learning Research*, 22(1):7479–7551, 2021.

[27] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[28] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

[29] Jean-Baptiste Cordonnier. Convex optimization using sparsified stochastic gradient descent with memory. Technical report, 2018.

[30] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[31] Tim Dettmers. 8-bit approximations for parallelism in deep learning. *arXiv preprint arXiv:1511.04561*, 2015.

[32] 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.

[33] Samuel Horvóth, Chen-Yu Ho, Ludovit Horvath, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. In *Mathematical and Scientific Machine Learning*, pages 129–141. PMLR, 2022.

[34] Youjie Li, Jongse Park, Mohammad Alian, Yifan Yuan, Zheng Qu, Peitian Pan, Ren Wang, Alexander Schwing, Hadi Esmaeilzadeh, and Nam Sung Kim. A network-centric hardware/algorithm co-design to accelerate distributed training of deep neural networks. In *2018 51st Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 175–188. IEEE, 2018.

[35] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 30, 2017.

[36] Yue Yu, Jiaxiang Wu, and Junzhou Huang. Exploring fast and communication-efficient algorithms in large-scale distributed networks. *arXiv preprint arXiv:1901.08924*, 2019.

[37] Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34:362–377, 2021.

[38] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

[39] Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

[40] Haobo Sun, Yingxia Shao, Jiawei Jiang, Bin Cui, Kai Lei, Yu Xu, and Jiang Wang. Sparse gradient compression for distributed sgd. In *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part II*, pages 139–155. Springer, 2019.

[41] Yusuke Tsuzuku, Hiroto Imachi, and Takuya Akiba. Variance-based gradient compression for efficient distributed deep learning. *arXiv preprint arXiv:1802.06058*, 2018.

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

[43] Nikoli Dryden, Tim Moon, Sam Ade Jacobs, and Brian Van Essen. Communication quantization for data-parallel training of deep neural networks. In *2016 2nd Workshop on Machine Learning in HPC Environments (MLHPC)*, pages 1–8. IEEE, 2016.

[44] Hyeontaek Lim, David G Andersen, and Michael Kaminsky. 3lc: Lightweight and effective traffic compression for distributed machine learning. *Proceedings of Machine Learning and Systems*, 1:53–64, 2019.

[45] Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander Schwing, Murali Annavaram, and Salman Avestimehr. Gradiveq: Vector quantization for bandwidth-efficient gradient aggregation in distributed cnn training. *Advances in Neural Information Processing Systems*, 31, 2018.

[46] Minsik Cho, Vinod Muthusamy, Brad Nemanich, and Ruchir Puri. Gradzip: Gradient compression using alternating matrix factorization for large-scale deep learning. In *NeurIPS*. 2019.

[47] Rajarshi Saha, Stefano Rini, Milind Rao, and Andrea J. Goldsmith. Decentralized optimization over noisy, rate-constrained networks: Achieving consensus by communicating differences. *IEEE Journal on Selected Areas in Communications*, 40(2):449–467, 2022.

[48] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed updates. *The Journal of Machine Learning Research*, 21(1):9613–9648, 2020.

[49] 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*, pages 3252–3261. PMLR, 2019.

[50] GW Stewart and JH Miller. Methods of simultaneous iteration for calculating eigenvectors of matrices. *Topics in Numerical Analysis II*, 2, 1975.

[51] Matteo Pagliardini. GPT-2 modular codebase implementation. *https://github.com/epfml/llm-baselines*, 2023.

[52] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR, 2020.

[53] Farzin Haddadpour, Belhal Karimi, Ping Li, and Xiaoyun Li. Fedsketch: Communication-efficient and private federated learning via sketching. *CoRR*, abs/2008.04975, 2020.

[54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

Organization. The appendix is organized as follows:

- App. A provides the related work.
- App. B contains the requisite material for error feedback and SGD convergence analysis.
- App. C details the important technical lemmas needed for the theoretical convergence of LASER.
- App. D provides the proof for Theorem 1 whereasApp. E contains the proofs of all technical lemmas.
- App. F provides additional details about the noisy channel and Algorithm 1.
- App. G contains additional experimental details and results.
- App. H details further analysis of LASER through the lens of different power policies, communication cost, complexity, etc.

A Related work

(i) Compression schemes with noiseless communication. Assuming a noiseless bit pipe from clients to the server, quantization methods [31, 32, 33, 34, 35, 36, 37] quantize each coordinate and send as fewer bits as possible. Sparsification techniques [38, 39, 40, 41, 42] send a reduced number of coordinates, based on criteria such as Top/Random-K, as opposed to sending the full gradient directly. Hybrid methods [43, 44] combine both. Rank compression methods [45, 46, 11] spectrally decompose gradient matrix (often via SVD) and transmit these factors. Since SVD is computationally prohibitive, we rely on the state-of-the-art light-weight compressor PowerSGD [12]. **(ii) Compression schemes for noisy channels.** The main idea here is to enable over-the-air-aggregation of gradients via the superposition nature of wireless channels [21] thus reducing the communication latency and bandwidth. The popular A-DSGD [15] relies on Top-K sparsification and random sketching. However, being memory intensive, A-DSGD is restricted to MNIST with 1-layer NN and doesn't scale beyond. [13] propose an analog-gradient-aggregation scheme but it is limited to shallow neural networks. [17] design a digital quantizer for training over Gaussian MAC channels. **(iii) Power laws.** In the absence of explicit power constraints, [20] show that $\mathcal{O}(1/t^2)$ noise-decay ensures the standard $1/T$ convergence rate for noisy FED-AVG whereas [47] propose a $t^{0.8}$ increase in SNR for the decentralized setup.

B Error feedback and SGD convergence toolbox

In this section we briefly recall the main techniques for the convergence analysis of SGD with error feedback (EF-SGD) from [48]. We consider $k=1$ clients with a compressor $\mathcal{C}_r(\cdot)$ and without any channel communication noise $\mathcal{Z}_P$(Sec. 2):

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) \\ \boldsymbol{e}_{t+1} &= (\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) - \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t). \end{aligned} \tag{EF-SGD}$$

Now we define the virtual iterates $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ which are helpful for the convergence analysis:

$$\widetilde{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \boldsymbol{e}_t. \tag{5}$$

Hence $\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{e}_t - \gamma_t \boldsymbol{g}_t = \widetilde{\boldsymbol{\theta}}_t - \gamma_t \boldsymbol{g}_t$. Now we detail the assumptions.

Assumption 1. *The objective $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is differentiable and μ-quasi-convex for a constant $\mu \geq 0$ with respect to $\boldsymbol{\theta}_\star$, i.e. $f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}_\star) + \frac{\mu}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|^2 \leq \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}_\star\rangle,\ \forall\, \boldsymbol{\theta} \in \mathbb{R}^{m\times n}$.*

Assumption 2. *f is L-smooth for some $L > 0$, i.e. $f(\boldsymbol{\theta}') \leq f(\boldsymbol{\theta}) + \langle \nabla f(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta}\rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|^2,\ \forall\, \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^{m\times n}$.*

Assumption 3. *For any $\boldsymbol{\theta}$, a gradient oracle $\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \nabla f(\boldsymbol{\theta}) + \boldsymbol{\xi}$, and conditionally independent noise $\boldsymbol{\xi}$, there exist scalars $(M, \sigma^2) \geq 0$ such that $\mathbb{E}\left[\boldsymbol{\xi}|\boldsymbol{\theta}\right] = 0,\ \mathbb{E}[\|\boldsymbol{\xi}\|^2|\boldsymbol{\theta}] \leq M\|\nabla f(\boldsymbol{\theta})\|^2 + \sigma^2$.*

Assumption 4. *The compressor $\mathcal{C}_r(\cdot)$ satisifes the δ_r-compression property: there exists a $\delta_r \in [0,1]$ such that $\mathbb{E}_{\mathcal{C}_r}\|\mathcal{C}_r(\boldsymbol{M}) - \boldsymbol{M}\|^2 \leq (1-\delta_r)\|\boldsymbol{M}\|^2,\ \forall\, \boldsymbol{M} \in \mathbb{R}^{m\times n}$.*

δ_r-compression is a standard assumption in the convergence analysis of Error Feedback SGD (EF-SGD) [48]. It ensures that the norm of the feedback memory remains bounded. We make the following assumption on the influence factor λ_{LASER}, which ensures that the overall composition of the channel and compressor mappings, $\mathcal{Z}_P(\mathcal{C}_r(\cdot))$, still behaves nicely.

Assumption 5. *The channel influence factor λ_{LASER} satisfies $\lambda_{\text{LASER}} \leq 1/(10(2/\delta_r + M))$.*

We note that a similar assumption is needed for convergence even in the hypothetical ideal scenario when the clients have access to the channel output (§ C.2), which we do not have. This bound can be roughly interpreted as $\lambda_{\text{LASER}} = \mathcal{O}(\delta_r)$.

First we consider the case when f is quasi-convex followed by the non-convex setting. In all the results below, we assume that the objective f is L-smooth, gradient oracle $\boldsymbol{g}$ has (M, σ^2)-bounded noise, and that $\mathcal{C}_r(\cdot)$ satisfies the δ_r compression property (Assumptions 2, 3, and 4).

f is quasi-convex:

The following lemma gives a handle on the gap to optimality $\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2$.

Lemma 1 ([48, Lemma 8]). *Let $\{\boldsymbol{\theta}_t, \boldsymbol{e}_t\}_{t\geq 0}$ be defined as in EF-SGD. Assume that f is μ-quasi convex for some $\mu \geq 0$. If $\gamma_t \leq \frac{1}{4L(1+M)}$ for all $t \geq 0$, then for $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ defined in Eq.* (5),

$$\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - \frac{\gamma_t}{2}\mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star) + \gamma_t^2\sigma^2 + 3L\gamma_t\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 . \tag{6}$$

The following lemma bounds the squared norm of the error, i.e. $\mathbb{E}\|\boldsymbol{e}_t\|^2$, appearing in Eq. (6). Recall that a positive sequence $\{a_t\}_{t\geq 0}$ is τ-slow decreasing for parameter $\tau \geq 1$ if $a_{t+1} \leq a_t$ and $a_{t+1}(1 + 1/2\tau) \geq a_t$. The sequence $\{a_t\}_{t\geq 0}$ is τ-slow increasing if $\{a_t^{-1}\}_{t\geq 0}$ is τ-slow decreasing [48, Definition 10].

Lemma 2 ([48, Lemma 22]). *Let $\boldsymbol{e}_t$ be as in* (EF-SGD) *for a δ_r-approximate compressor $\mathcal{C}_r$ and stepsizes $\{\gamma_t\}_{t\geq 0}$ with $\gamma_{t+1} \leq \frac{1}{10L(2/\delta_r + M)}$, $\forall t \geq 0$ and $\{\gamma_t^2\}_{t\geq 0}$ $\frac{2}{\delta_r}$-slow decaying. Then*

$$\mathbb{E}\left[3L\|\boldsymbol{e}_{t+1}\|^2\right] \leq \frac{\delta_r}{64L}\sum_{i=0}^{t}\left(1 - \frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \gamma_t\sigma^2 . \tag{7}$$

Furthermore, for any $\frac{4}{\delta_r}$-slow increasing non-negative sequence $\{w_t\}_{t\geq 0}$ it holds:

$$3L\sum_{t=0}^{T} w_t\mathbb{E}\|\boldsymbol{e}_t\|^2 \leq \frac{1}{8L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2\right) + \sigma^2\sum_{t=0}^{T} w_t\gamma_t .$$

The following result controls the summations of the optimality gap that appear when combining Lemma 1 and Lemma 2.

Lemma 3 ([48, Lemma 13]). *For every non-negative sequence $\{r_t\}_{t\geq 0}$ and any parameters $d \geq a > 0$, $c \geq 0$, $T \geq 0$, there exists a constant $\gamma \leq \frac{1}{d}$, such that for constant stepsizes $\{\gamma_t = \gamma\}_{t\geq 0}$ and weights $w_t := (1 - a\gamma)^{-(t+1)}$ it holds*

$$\Psi_T := \frac{1}{W_T}\sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1 - a\gamma_t\right)r_t - \frac{w_t}{\gamma_t}r_{t+1} + c\gamma_t w_t\right) = \widetilde{\mathcal{O}}\left(dr_0\exp\left[-\frac{aT}{d}\right] + \frac{c}{aT}\right) .$$

Combining the above lemmas, we obtain the following result for the convergence rate of EF-SGD.

Theorem 2 ([48, Theorem 22]). *Let $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ denote the iterates of the error compensated stochastic gradient descent* (EF-SGD) *with constant stepsize $\{\gamma_t = \gamma\}_{t\geq 0}$ and with a δ_r-approximate compressor on a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ under Assumptions 2 and 3. Then, if f*

- *satisfies Assumption 1 for $\mu > 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(2/\delta_r+M)}$ (chosen as in Lemma 3) such that*

  where the output $\boldsymbol{\theta}_{\text{out}} \in \{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$ is chosen to be $\boldsymbol{\theta}_t$ with probability proportional to $(1-\mu\gamma/2)^{-t}$.
- *satisfies Assumption 1 for $\mu = 0$, then there exists a stepsize $\gamma \leq \frac{1}{10L(2/\delta_r+M)}$ (chosen as in Lemma 3) such that*

$$\mathbb{E}f(\boldsymbol{\theta}_{\text{out}}) - f_* = \mathcal{O}\left(\frac{L(1/\delta_r + M)\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|^2}{T} + \frac{\sigma\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|}{\sqrt{T}}\right),$$

  where the output $\boldsymbol{\theta}_{\text{out}} \in \{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$ is chosen uniformly at random from the iterates $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$.

<u>f **is non-convex:**</u>

Now we consider the case where f is an arbitrary non-convex function. The above set of results extend in a similar fashion to this setting too as described below:

Lemma 4 ([48, Lemma 9])**.** *Let $\{\boldsymbol{\theta}_t, \boldsymbol{e}_t\}_{t\geq 0}$ be defined as in EF-SGD. If $\gamma_t \leq \frac{1}{2L(1+M)}$ for all $t \geq 0$, then for $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ defined in Eq.* (5)*,*

$$\mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_{t+1})] \leq \mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_t)] - \frac{\gamma_t}{4}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2}{2} + \frac{\gamma_t L^2}{2}\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2\,. \tag{8}$$

Lemma 5 ([48, Lemma 22])**.** *Let $\boldsymbol{e}_t$ be as in* (EF-SGD) *for a δ_r-approximate compressor $\mathcal{C}_r$ and stepsizes $\{\gamma_t\}_{t\geq 0}$ with $\gamma_{t+1} \leq \frac{1}{10L(2/\delta_r+M)}$, $\forall t \geq 0$ and $\{\gamma_t^2\}_{t\geq 0}$ $\frac{2}{\delta_r}$-slow decaying. Then*

$$\mathbb{E}\left[3L\|\boldsymbol{e}_{t+1}\|^2\right] \leq \frac{\delta_r}{64L}\sum_{i=0}^{t}\left(1 - \frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \gamma_t\sigma^2\,. \tag{9}$$

Furthermore, for any $\frac{4}{\delta_r}$-slow increasing non-negative sequence $\{w_t\}_{t\geq 0}$ it holds:

$$3L\sum_{t=0}^{T} w_t\mathbb{E}\|\boldsymbol{e}_t\|^2 \leq \frac{1}{8L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \sigma^2\sum_{t=0}^{T} w_t\gamma_t\,.$$

Lemma 6 ([48, Lemma 14])**.** *For every non-negative sequence $\{r_t\}_{t\geq 0}$ and any parameters $d \geq 0$, $c \geq 0$, $T \geq 0$, there exists a constant $\gamma \leq \frac{1}{d}$, such that for constant stepsizes $\{\gamma_t = \gamma\}_{t\geq 0}$ it holds:*

$$\Psi_T := \frac{1}{T+1}\sum_{t=0}^{T}\left(\frac{r_t}{\gamma_t} - \frac{r_{t+1}}{\gamma_t} + c\gamma_t\right) \leq \frac{dr_0}{T+1} + \frac{2\sqrt{cr_0}}{\sqrt{T+1}}\,.$$

Now we have the final convergence result for the non-convex setting.

Theorem 3 ([48, Theorem 22])**.** *Let $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ denote the iterates of the error compensated stochastic gradient descent* (EF-SGD) *with constant stepsize $\{\gamma_t = \gamma\}_{t\geq 0}$ and with a δ_r-approximate compressor on a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ under Assumptions 2 and 3. Then, if f is an arbitrary non-convex function, there exists a stepsize $\gamma \leq \frac{1}{10L(1/\delta_r+M)}$ (chosen as in Lemma 6), such that*

$$\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{\text{out}})\|^2 = \mathcal{O}\left(\frac{L(1/\delta_r + M)(f(\boldsymbol{\theta}_0) - f_\star)}{T} + \sigma\sqrt{\frac{L(f(\boldsymbol{\theta}_0) - f_\star)}{T}}\right).$$

where the output $\boldsymbol{\theta}_{\text{out}} \in \{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$ is chosen uniformly at random from the iterates $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$.

C Technical lemmas for LASER convergence

Towards the convergence analysis of LASER for $k = 1$, we rewrite the Algorithm 1 succinctly as:

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)\right) \\ \boldsymbol{e}_{t+1} &= (\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) - \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t), \end{aligned} \tag{LASER}$$

where the channel corrupted gradient approximation $\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\cdot)$ is given by

$$\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\underbrace{\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)}_{=\boldsymbol{P}\boldsymbol{Q}^\top}) \triangleq \sum_{i=1}^{r} \left(\boldsymbol{p}_i + \frac{\|\boldsymbol{p}_i\|}{\sqrt{\alpha_i}} \cdot \boldsymbol{Z}_m^{(i)}\right)\left(\boldsymbol{q}_i + \frac{\|\boldsymbol{q}_i\|}{\sqrt{\beta_i}} \cdot \boldsymbol{Z}_n^{(i)}\right)^\top, \tag{10}$$

and $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^r$ and $\boldsymbol{\beta} = (\beta_i)_{i=1}^r$ are appropriate power allocations to transmit the respective left and right factors $\boldsymbol{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_r] \in \mathbb{R}^{m\times r}$ and $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r] \in \mathbb{R}^{n\times r}$ for the decomposition $\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) = \boldsymbol{P}\boldsymbol{Q}^\top$. $\boldsymbol{Z}_m^{(i)} \in \mathbb{R}^m$ and $\boldsymbol{Z}_n^{(i)} \in \mathbb{R}^n$ denote the independent channel noises for each factor $i \in [r]$.

Thus we observe from LASER that it has an additional channel corruption in the form of $\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\cdot)$ as compared to the EF-SGD. Now in the remainder of this section, we explain how to choose the power allocation $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (App. C.1), how to control the influence of the channel $\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\cdot)$ on the convergence of LASER (App. C.2), and utilize these results to establish technical lemmas along the lines of App. B for LASER (App. C.3).

C.1 Power allocation

In this section, we introduce the key technical lemmas about power allocation that are crucial for the theoretical results. We start with the rank one case.

Lemma 7 (Rank-1 power allocation). *For a power $P > 0$ and $m, n \in \mathbb{N}$ with $m \le n$, define the function $f_P : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ as*

$$f_P(\alpha, \beta) \triangleq \left(1 + \frac{m}{\alpha}\right)\left(1 + \frac{n}{\beta}\right),$$

and the constraint set $S_P \triangleq \{(\alpha, \beta) : \alpha \ge 0, \beta \ge 0, \alpha + \beta = P\}$. Then for the minimizer $(\alpha^\star, \beta^\star) = \mathrm{argmin}_{(\alpha,\beta)\in S_P} f_P(\alpha, \beta)$, we have

$$f_P(\alpha^\star, \beta^\star) \le 1 + \frac{4}{m\,\mathrm{SNR}}\left(1 + \frac{1}{n\,\mathrm{SNR}}\right), \quad \mathrm{SNR} \triangleq \frac{P}{mn}.$$

Further the minimizer is given by

$$\alpha^\star = \begin{cases} \sqrt{1 + \frac{P}{n}}\left(\frac{\sqrt{1+\frac{P}{m}} - \sqrt{1+\frac{P}{n}}}{\frac{1}{m} - \frac{1}{n}}\right), & m \ne n \\ P/2, & m = n \end{cases}$$
$$\beta^\star = P - \alpha^\star.$$

Lemma 8 (Rank-r power allocation). *For a power $P > 0$, $m, n, r \in \mathbb{N}$ with $m \le n$, and positive scalars $\kappa_1, \ldots, \kappa_r > 0$ with $\sum_i \kappa_i = 1$, define the function $f_P : (\mathbb{R}_+)^r \times (\mathbb{R}_+)^r \to \mathbb{R}_+$ as*

$$f_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) \triangleq \sum_{i=1}^{r} \kappa_i \left(1 + \frac{m}{\alpha_i}\right)\left(1 + \frac{n}{\beta_i}\right), \quad \boldsymbol{\alpha} = (\alpha_i)_{i=1}^r, \boldsymbol{\beta} = (\beta_i)_{i=1}^r,$$

and the constraint set $S_P \triangleq \{\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \boldsymbol{\alpha} \ge 0, \boldsymbol{\beta} \ge 0, \sum_i(\alpha_i + \beta_i) = P\}$. Then there exists a power allocation scheme $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \in S_P$ such that

$$\min_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in S_P} f_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) \le f_P(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \le 1 + \frac{4}{(m/r)\,\mathrm{SNR}}\left(1 + \frac{1}{(n/r)\,\mathrm{SNR}}\right),$$

where $\mathrm{SNR} \triangleq \frac{P}{mn}$. *Further* $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ *is given by*

$$\alpha_i^\star = \begin{cases} \sqrt{1+\frac{P_i}{n}} \left(\frac{\sqrt{1+\frac{P_i}{m}} - \sqrt{1+\frac{P_i}{n}}}{\frac{1}{m}-\frac{1}{n}} \right), & m \neq n \\ P_i/2, & m = n \end{cases}$$
$$\beta_i^\star = P_i - \alpha_i^\star,$$
$$P_i = P \left(\frac{\sqrt{\kappa_i}}{\sum_j \sqrt{\kappa_j}} \right).$$

Remark 1. In other words, we first divide the power P proportional to $\sqrt{\kappa_i}$ for each $i \in [r]$ and further allocate this P_i amongst $\alpha_i^\star$ and $\beta_i^\star$ as per the optimal rank one allocation scheme in Lemma 7.

C.2 Channel influence factor

In this section we establish the bounds for the channel influence defined in Eq. (3) for both Z-SGD and LASER. This helps us give a handle to control the second moment of the gradient corrupted by channel noise.

Lemma 9 (Channel influence on Z-SGD). *For the* Z-SGD *algorithm that sends the uncompressed gradients directly over the noisy channel with power constraint* P*, we have*

$$\lambda_{\text{Z-SGD}} = \frac{1}{\mathrm{SNR}}, \tag{11}$$

where $\mathrm{SNR} = \frac{P}{mn}$.

Lemma 10. *For the* LASER *algorithm with the optimal power allocation* $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ *(chosen as in Lemma 8), we have*

$$\lambda_{\text{LASER}} \leq \frac{4}{(m/r)\,\mathrm{SNR}} \left(1 + \frac{1}{(n/r)\,\mathrm{SNR}}\right), \tag{12}$$

where $\mathrm{SNR} = \frac{P}{mn}$.

Remark 2. Note that for the optimal power allocation via Lemma 8, we need the positive scalars $\kappa_1, \ldots, \kappa_r$. In the context of LASER, we will later see in the proof in App. E that $\kappa_i \propto \|\boldsymbol{p}_i\|^2$.

Thus Lemma 9 and Lemma 10 establish that

$$\lambda_{\text{LASER}} \leq \frac{4}{(m/r)\,\mathrm{SNR}} \left(1 + \frac{1}{(n/r)\,\mathrm{SNR}}\right) \ll \frac{1}{\mathrm{SNR}} = \lambda_{\text{Z-SGD}}.$$

In the low-rank [12] and constant-order SNR regime where $r = \mathcal{O}(1)$ and $\mathrm{SNR} = \Omega(1)$, we observe that λ_{LASER} is roughly $\mathcal{O}(m)$ times smaller than $\lambda_{\text{Z-SGD}}$.

Note on assumption between λ_{LASER} and δ_r. Recall from LASER that the local memory $\boldsymbol{e}_t$ has only access to the compressed gradients and not the channel output. In an hypothetical scenario, where it has access to the same, it follows that $\mathbb{E}_{\boldsymbol{Z}} \|\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\mathcal{C}_r(\boldsymbol{M})) - \boldsymbol{M}\|^2 \leq (1 - (\delta_r - \lambda_{\text{LASER}}))\|\boldsymbol{M}\|^2$. Hence for the compression property in this ideal scenario, we need $\lambda_{\text{LASER}} \leq \delta_r$.

C.3 Optimality gap and error bounds for LASER iterates

In this section, we characterize the gap to the optimality and the error norm for the LASER iterates $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ (similar to Lemmas 1, 2, 2 and 5 for EF-SGD). Towards the same, first we define the virtual iterates $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ as follows:

$$\widetilde{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \boldsymbol{e}_t\,. \tag{13}$$

Thus,

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_{t+1} - \boldsymbol{e}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \gamma_t \boldsymbol{g}_t + \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)\right). \tag{14}$$

The following lemma controls the optimality gap $\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2$ when f is quasi-convex.

Lemma 11 (Descent for quasi-convex). *Let $\{\boldsymbol{\theta}_t, \boldsymbol{e}_t\}_{t\geq 0}$ be defined as in* LASER. *Assume that f is μ-quasi convex for some $\mu \geq 0$ and that Assumptions 2 and 3 hold. If $\gamma_t \leq \frac{1}{4L(1+M)}\left(\frac{1-2\lambda_{\text{LASER}}}{1+\lambda_{\text{LASER}}}\right)$ for all $t \geq 0$, then for $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ defined in Eq.* (13)*,*

$$\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - \frac{\gamma_t}{2}\mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star) + \gamma_t^2\sigma^2(1+\lambda_{\text{LASER}}) + (3L\gamma_t(1+\lambda_{\text{LASER}}) + \lambda_{\text{LASER}})\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2\,. \tag{15}$$

Notice that Lemma 11 is similar to Lemma 1 for noiseless EF-SGD except for an additional channel influence factor λ_{LASER}. The following result bounds the error norm.

Lemma 12 (Error control). *Let $\mathbf{e}_t$ be as in* (LASER) *for a δ_r-approximate compressor $\mathcal{C}_r$ and stepsizes $\{\gamma_t\}_{t\geq 0}$ with $\gamma_t \leq \frac{1}{10L(2/\delta_r+M)(1+\lambda_{\text{LASER}})}$, $\forall t \geq 0$ and $\{\gamma_t^2\}_{t\geq 0}$ $\frac{2}{\delta_r}$-slow decaying. Further suppose that Assumption 5 holds. Then*

$$\left(3L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\mathbb{E}\|\boldsymbol{e}_{t+1}\|^2 \leq \frac{\delta_r}{32L}\sum_{i=0}^{t}\left(1 - \frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \gamma_t\sigma^2(1+\lambda_{\text{LASER}})\,. \tag{16}$$

Furthermore, for any $\frac{4}{\delta_r}$-slow increasing non-negative sequence $\{w_t\}_{t\geq 0}$ it holds:

$$\left(3L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\sum_{t=0}^{T} w_t\mathbb{E}\|\boldsymbol{e}_t\|^2 \leq \frac{1}{6L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2\right) + \sigma^2(1+\lambda_{\text{LASER}})\sum_{t=0}^{T} w_t\gamma_t\,. \tag{17}$$

The following lemma establishes the progress in the descent for non-convex case.

Lemma 13 (Descent for non-convex). *Let $\{\boldsymbol{\theta}_t, \boldsymbol{e}_t\}_{t\geq 0}$ be defined as in* LASER *and that Assumptions 2 and 3 hold. If $\gamma_t \leq \frac{1}{4L(1+M)(1+\lambda_{\text{LASER}})}$ for all $t \geq 0$, then for $\{\widetilde{\boldsymbol{\theta}}_t\}_{t\geq 0}$ defined in Eq.* (13)*,*

$$\mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_{t+1})] \leq \mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_t)] - \frac{\gamma_t}{4}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2(1+\lambda_{\text{LASER}})}{2} + \mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2\left(\frac{L^2\gamma_t}{2} + L\lambda_{\text{LASER}}\right)\,. \tag{18}$$

D Proof of Theorem 1

Proof. We prove the bounds in (i) and (ii) when f is quasi-convex, (iii) when f is an arbitrary non-convex function, and (iv) for Z-SGD.

(i), (ii) f is μ-quasi-convex: Observe that the assumptions of Theorem 1 automatically satisfy the conditions of Lemma 11. Denoting $r_t \triangleq \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2$ and $s_t \triangleq \mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star)$, for any $w_t > 0$ we obtain

$$\frac{w_t}{2}s_t \overset{(15)}{\leq} \frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right)r_t - \frac{w_t}{\gamma_t}r_{t+1} + \gamma_t w_t\sigma^2(1+\lambda_{\text{LASER}}) + 3w_t(L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t})\mathbb{E}\|\boldsymbol{e}_t\|^2\,.$$

Taking summation on both sides and invoking Lemma 2 (assumption on w_t verified below),

$$\sum_{t=0}^{T}\frac{w_t}{2}s_t \overset{(17)}{\leq} \sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right)r_t - \frac{w_t}{\gamma_t}r_{t+1} + 2\gamma_t w_t\sigma^2(1+\lambda_{\text{LASER}})\right) + \frac{1}{6L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2\right)\,.$$

Since f is L-smooth, we have $\|\nabla f(\boldsymbol{\theta}_t)\|^2 \leq 2L(f(\boldsymbol{\theta}_t) - f_\star)$. Now rewriting the above inequality, we have

$$\frac{1}{6}\sum_{t=0}^{T} w_t s_t \leq \sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1-\frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + 2\gamma_t w_t \sigma^2 (1+\lambda_{\text{LASER}})\right).$$

Substituting $W_T \triangleq \sum_{t=0}^{T} w_t$,

$$\frac{1}{W_T}\sum_{t=0}^{T} w_t s_t \leq \frac{6}{W_T}\sum_{t=0}^{T}\left(\frac{w_t}{\gamma_t}\left(1-\frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + 2\gamma_t w_t \sigma^2 (1+\lambda_{\text{LASER}})\right) =: \Xi_T \,.$$

Now it remains to derive the estimate for Ξ_T. Towards this, (i) if $\mu > 0$ and with constant stepsize $\gamma_t = \gamma \leq \frac{1}{10L(\frac{2}{\delta_r}+M)(1+\lambda_{\text{LASER}})}$, we observe that $(1-\frac{\mu\gamma}{2}) \geq \left(1-\frac{\delta_r}{16}\right)$ and by [48, Example 1], the weights $w_t = \left(1-\frac{\mu\gamma}{2}\right)^{-(t+1)}$ are 2τ-slow increasing with $\tau = \frac{2}{\delta_r}$. Hence the claim in (i) follows by applying Lemma 3 and observing that the sampling probablity to choose $\boldsymbol{\theta}_{\text{out}}$ from $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$ is same as w_t.

For (ii) with constant stepsize and $\mu = 0$, we apply Lemma 6 by setting the weights $w_t = 1$.

<u>**(iii) f is non-convex**</u> The proof in this case is very similar to that of the above. Denoting $r_t \triangleq 4\mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_t) - f_\star]$, $s_t \triangleq \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2$, $c = 4L\sigma^2(1+\lambda_{\text{LASER}})$, and $w_t = 1$, we have from Lemma 13 that

$$\frac{s_t}{4} \overset{(18)}{\leq} \frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8} + L\left(\frac{L}{2} + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\mathbb{E}\|\boldsymbol{e}_t\|^2.$$

Since $\frac{L}{2} \leq 3L(1+\lambda_{\text{LASER}})$, multiplying both sides of the above inequality by w_t and taking summation, we obtain

$$\frac{1}{4W_T}\sum_{t=0}^{T} w_t s_t \overset{(17)}{\leq} \frac{1}{W_T}\sum_{t=0}^{T} w_t\left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8}\right) + \frac{L}{W_T}\left(\sum_{t=0}^{T}\frac{w_t s_t}{6L} + \frac{c w_t \gamma_t}{4L}\right),$$

which upon rearranging gives

$$\frac{1}{W_T}\sum_{t=0}^{T} w_t s_t \leq \frac{12}{W_T}\sum_{t=0}^{T} w_t\left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{3\gamma_t c}{8}\right).$$

Now invoking Lemma 6 yields the final result in (iii).

<u>**Z-SGD:**</u> Recall from Z-SGD that the iterates $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ are given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t\, \mathcal{Z}_P(\boldsymbol{g}_t).$$

Thus Z-SGD can be thought of as a special case of EF-SGD with no compression, i.e. $\delta_r = 1$, and hence we can utilize the same convergence tools. It remains to estimate the first and second moments of the stochastic gradient $\mathcal{Z}_P(\boldsymbol{g}_t)$. Recall from the definition of $\mathcal{Z}_P$ in the noisy channel that $\mathcal{Z}_P(\boldsymbol{g}_t) = \boldsymbol{g}_t + \frac{\|\boldsymbol{g}_t\|}{\sqrt{P}}\boldsymbol{Z}_t$, where $\boldsymbol{Z}_t$ is a zero-mean independent channel noise, and from Assumption 3 that $\boldsymbol{g}_t = \nabla f(\boldsymbol{\theta}_t) + \boldsymbol{\xi}_t$ with a (M, σ^2)-bounded noise $\boldsymbol{\xi}_t$. Hence

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}_P(\boldsymbol{g}_t)|\boldsymbol{\theta}_t\right] &= \mathbb{E}[\boldsymbol{g}_t|\boldsymbol{\theta}_t] = \nabla f(\boldsymbol{\theta}_t),\\
\mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \boldsymbol{g}_t + \boldsymbol{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right]\\
&= \mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \boldsymbol{g}_t\|^2|\boldsymbol{\theta}_t\right] + \mathbb{E}\left[\|\boldsymbol{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right]\\
&\overset{3}{=} \mathbb{E}\left[\lambda_{\text{Z-SGD}}\|\boldsymbol{g}_t\|^2|\boldsymbol{\theta}_t\right] + \mathbb{E}\|\boldsymbol{\xi}_t\|^2\\
&= \lambda_{\text{Z-SGD}}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + (1+\lambda_{\text{Z-SGD}})\mathbb{E}\|\boldsymbol{\xi}_t\|^2\\
&\leq (M+1)(1+\lambda_{\text{Z-SGD}})\|\nabla f(\boldsymbol{\theta}_t)\|^2 + (1+\lambda_{\text{Z-SGD}})\sigma^2.
\end{aligned}$$

Thus Z-SGD satifies the $(\widetilde{M}, \widetilde{\sigma}^2)$-bounded noise condition in Assumption 3 with $\widetilde{M} = (M+1)(1+\lambda_{\text{Z-SGD}})$ and $\widetilde{\sigma}^2 = (1+\lambda_{\text{Z-SGD}})\sigma^2$. Thus the claim (iv) follows from applying Theorem 2 and Theorem 3 with the constants $\delta_r \to 1, M \to \widetilde{M}, \sigma^2 \to \widetilde{\sigma}^2$.

Finally, Lemma 9 and Lemma 10 establish the relation between the channel influence factors $\lambda_{\text{Z-SGD}}$ and λ_{LASER}.

$\square$

E Proof of technical lemmas

E.1 Proof of Lemma 7

Proof. Since $\log(\cdot)$ is a monotonic function, minimizing $f_P(\alpha, \beta)$ over $S_P = \{(\alpha, \beta) : \alpha \geq 0, \beta \geq 0, \alpha + \beta = P\}$ is equivalent to minimizing $\log f_P(\alpha, \beta) = \log\left(1 + \frac{m}{\alpha}\right) + \log\left(1 + \frac{n}{\beta}\right)$. Define the Lagrangian $L(\alpha, \beta, \lambda)$ as

$$L(\alpha, \beta, \lambda) \triangleq \log\left(1 + \frac{m}{\alpha}\right) + \log\left(1 + \frac{n}{\beta}\right) + \lambda(\alpha + \beta - P).$$

Letting $\nabla_\alpha L = \nabla_\beta L = 0$, we obtain that $\frac{m}{\alpha(m+\alpha)} = \frac{n}{\beta(n+\beta)}$. Now constraining $\alpha + \beta = P$, we obtain the following quadratic equation:

$$\alpha^2 \left(\frac{1}{m} - \frac{1}{n}\right) + 2\alpha \left(1 + \frac{P}{n}\right) - \left(\frac{P^2}{n} + P\right) = 0.$$

If $m = n$, the solution is given by $\alpha^\star = \beta^\star = P/2$. If $m \neq n$, the solution is given by

$$\begin{aligned} \alpha^\star &= \sqrt{1 + \frac{P}{n}} \left(\frac{\sqrt{1 + \frac{P}{m}} - \sqrt{1 + \frac{P}{n}}}{\frac{1}{m} - \frac{1}{n}}\right), \\ \beta^\star &= P - \alpha^\star. \end{aligned} \tag{19}$$

It is easy to verify that $(\alpha^\star, \beta^\star)$ is the unique minimizer to f_P since it's convex over S_P. Now it remains to show the upper bound for $f_P(\alpha^\star, \beta^\star)$. Without loss of generality, in the reminder of the proof we assume $m < n$ and denote $\alpha^\star$ by simply α. Rewriting the optimal α in Eq. (19) in terms of $\text{SNR} = P/mn$, we obtain

$$\frac{\alpha}{mn} = \frac{\sqrt{(1 + n\,\text{SNR})(1 + m\,\text{SNR})} - (1 + m\,\text{SNR})}{n - m}. \tag{20}$$

Now substituting this α and corresponding β in $f_P(\alpha, \beta) = \left(1 + \frac{m}{\alpha}\right)\left(1 + \frac{n}{\beta}\right)$ and rearranging the terms, we get

$$\begin{aligned} f_P(\alpha, \beta) &= 1 + \frac{1}{\text{SNR}} \left(\frac{n - m}{mn}\right) \left(\frac{1}{1 - \frac{2\alpha}{mn\,\text{SNR}}}\right) \\ &= 1 + \frac{1}{n\,\text{SNR}} \left(\frac{\frac{n}{m} - 1}{1 - \frac{2\alpha}{mn\,\text{SNR}}}\right). \end{aligned}$$

Let $\gamma \triangleq \frac{m}{n} < 1$. Now we study the behavior of α in Eq. (20) as a function of γ. In particular, define $g(\gamma) \triangleq \sqrt{1 + n\,\text{SNR}}\sqrt{1 + n\gamma\,\text{SNR}}$. Observe that $g(1) = 1 + n\,\text{SNR}$ and $g'(1) = \frac{n\,\text{SNR}}{2}$. Rewriting Eq. (20) as a function of γ, we get

$$\begin{aligned} \frac{\alpha}{mn} &= \frac{g(\gamma) - (1 + n\gamma\,\text{SNR})}{n(1 - \gamma)} \\ &= \frac{g(1) + g'(1)(\gamma - 1) - (1 + n\gamma\,\text{SNR}) + \frac{g''}{2}(\gamma - 1)^2 + \frac{g'''}{3!}(\gamma - 1)^3 + \dots}{n(1 - \gamma)} \\ &= \frac{\text{SNR}}{2} + \frac{1}{n}\left(\frac{g''}{2}(1 - \gamma) - \frac{g'''}{3!}(1 - \gamma)^2 + \dots\right). \end{aligned}$$

Utilizing the fact that $g''(1) = \frac{-1}{4}\,\frac{n^2\mathrm{SNR}^2}{1+n\mathrm{SNR}}$, $g''(1) = \frac{3}{8}\,\frac{n^3\mathrm{SNR}^3}{(1+n\mathrm{SNR})^2}$ and so forth, we obtain

$$\begin{aligned} 1 - \frac{2\alpha}{mn\,\mathrm{SNR}} &= \frac{2(1-\gamma)}{n\,\mathrm{SNR}}\left(\frac{1}{2}\,\frac{1}{4}\,\frac{n^2\,\mathrm{SNR}^2}{1+n\,\mathrm{SNR}} + \frac{1}{3!}\,\frac{3}{8}\,\frac{n^3\,\mathrm{SNR}^3}{(1+n\,\mathrm{SNR})^2}(1-\gamma) + \dots\right) \\ &\geq \frac{2(1-\gamma)}{n\,\mathrm{SNR}}\,\frac{1}{2}\,\frac{1}{4}\,\frac{n^2\,\mathrm{SNR}^2}{1+n\,\mathrm{SNR}} \\ &= \frac{(1-\gamma)}{4}\,\frac{n\,\mathrm{SNR}}{1+n\,\mathrm{SNR}}. \end{aligned}$$

Substituting this bound back in the experssion for f_P yields the final bound:

$$\begin{aligned} f_P(\alpha,\beta) &\leq 1 + \frac{4}{n\gamma\,\mathrm{SNR}}\left(1 + \frac{1}{n\,\mathrm{SNR}}\right) \\ &= 1 + \frac{4}{m\,\mathrm{SNR}}\left(1 + \frac{1}{n\,\mathrm{SNR}}\right). \end{aligned}$$

□

E.2 Proof of Lemma 8

Proof. To minimize $f_P(\boldsymbol{\alpha},\boldsymbol{\beta})$ over $S_P = \{\{(\boldsymbol{\alpha},\boldsymbol{\beta}) : \boldsymbol{\alpha} \geq 0, \boldsymbol{\beta} \geq 0, \sum_i (\alpha_i + \beta_i) = P\}$, we consider a slightly relaxed version that serves as an upper bound to this problem. In particular, first we divide the power P into $P_1, \dots, P_r$ such that $\sum_i P_i = P$ and $P_i \geq 0$. Then for each P_i we find the optimal α_i and β_i from rank-1 allocation scheme in Lemma 7 and compute the corresponding objective value. In the end, we find a tractable scheme for division of power P among $P_1, \dots, P_r$ minimizing this objective. Mathematically,

$$\begin{aligned} \min_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in S_P} f_P(\boldsymbol{\alpha},\boldsymbol{\beta}) &\leq \min_{\{\sum_i P_i = P\}}\ \min_{\{(\alpha_i,\beta_i):\alpha_i+\beta_i=P_i, i\in[r]\}} \sum_i \kappa_i \left(1 + \frac{m}{\alpha_i}\right)\left(1 + \frac{n}{\beta_i}\right) \\ &= \min_{\{\sum_i P_i = P\}} \sum_i \kappa_i \min_{(\alpha_i,\beta_i):\alpha_i+\beta_i=P_i} \left(1 + \frac{m}{\alpha_i}\right)\left(1 + \frac{n}{\beta_i}\right) \\ &\overset{(Lemma\ 7)}{\leq} \min_{\{\sum_i P_i = P\}} \sum_i \kappa_i \left(1 + \frac{4}{m\mathrm{SNR}_i}\left(1 + \frac{1}{n\mathrm{SNR}_i}\right)\right), \quad \mathrm{SNR}_i \triangleq \frac{P_i}{mn}, \\ &= \min_{\{\sum_i P_i = P\}} \left(1 + \frac{4}{m}\sum_i \frac{\kappa_i}{\mathrm{SNR}_i} + \frac{4}{mn}\sum_i \frac{\kappa_i}{\mathrm{SNR}_i^2}\right). \end{aligned}$$

Choosing $\mathrm{SNR}_i \propto \sqrt{\kappa_i}$, i.e. $\mathrm{SNR}_i = \mathrm{SNR}\,\frac{\sqrt{\kappa_i}}{\sum_j \sqrt{\kappa_j}}$, and substituting this allocation above, we obtain

$$\begin{aligned} \min_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in S_P} f_P(\boldsymbol{\alpha},\boldsymbol{\beta}) &\leq 1 + \frac{4}{m\,\mathrm{SNR}}\left(\sum_i \sqrt{\kappa_i}\right)^2 + \frac{4}{mn\,\mathrm{SNR}^2}\,R\left(\sum_i \sqrt{\kappa_i}\right)^2 \\ &\leq 1 + \frac{4}{(m/r)\,\mathrm{SNR}}\left(1 + \frac{4}{(n/r)\,\mathrm{SNR}}\right), \end{aligned}$$

where we used the inequality $\left(\sum_i \sqrt{\kappa_i}\right)^2 \leq r$ together with the fact that $\sum_i \kappa_i = 1$. □

E.3 Proof of Lemma 9

Proof. Recall from Z-SGD that the stochastic gradient reconstructed at the receiver after transmitting $\boldsymbol{g}$ is $\boldsymbol{y}_{\text{Z-SGD}}(\boldsymbol{g}) \triangleq \mathcal{Z}_P(\boldsymbol{g}) = \boldsymbol{g} + \frac{\|\boldsymbol{g}\|}{\sqrt{P}}\,\boldsymbol{Z}$, where $\boldsymbol{Z}$ is a zero-mean independent channel noise in $\mathbb{R}^{m\times n}$. Thus

$$\lambda_{\text{Z-SGD}} = \frac{1}{\|\boldsymbol{g}\|^2}\,\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{y}_{\text{Z-SGD}}(\boldsymbol{g}) - \boldsymbol{g}\|^2 = \frac{1}{\|\boldsymbol{g}\|^2}\,\frac{\|\boldsymbol{g}\|^2}{P}\,\mathbb{E}\|\boldsymbol{Z}\|^2 = \frac{mn}{P} = \frac{1}{\mathrm{SNR}}.$$

□

E.4 Proof of Lemma 10

Proof. In view of LASER, denote the error compensated gradient at time t as $\boldsymbol{M} = \boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t$ and its compression as $\boldsymbol{M}_r = \mathcal{C}_r(\boldsymbol{M}) = \sum_{i=1}^r \boldsymbol{p}_i \boldsymbol{q}_i^\top$ with orthogonal factors $\{\boldsymbol{p}_i\}$ and orthonormal $\{\boldsymbol{q}_i\}$ (without loss of generality). After transmitting these factors of $\boldsymbol{M}_r$ via the noisy channel, we obtain

$$\boldsymbol{y}_{\text{LASER}}(\boldsymbol{M}_r) = \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\boldsymbol{M}_r) = \sum_{i=1}^r \left(\boldsymbol{p}_i + \frac{\|\boldsymbol{p}_i\|}{\sqrt{\alpha_i}} \cdot \boldsymbol{Z}_m^{(i)}\right)\left(\boldsymbol{q}_i + \frac{\|\boldsymbol{q}_i\|}{\sqrt{\beta_i}} \cdot \boldsymbol{Z}_n^{(i)}\right)^\top .$$

Denote $\widetilde{\boldsymbol{p}}_i \triangleq \boldsymbol{p}_i + \frac{\|\boldsymbol{p}_i\|}{\sqrt{\alpha_i}} \cdot \boldsymbol{Z}_m^{(i)}$, $\widetilde{\boldsymbol{q}}_i \triangleq \boldsymbol{q}_i + \frac{\|\boldsymbol{q}_i\|}{\sqrt{\beta_i}} \cdot \boldsymbol{Z}_n^{(i)}$, and $\boldsymbol{Z} = (\boldsymbol{Z}_m^{(i)}, \boldsymbol{Z}_n^{(i)})_{i=1}^r$. We observe that $\mathbb{E}_{\boldsymbol{Z}}[\boldsymbol{y}_{\text{LASER}}(\boldsymbol{M}_r)] = \boldsymbol{M}_r$. Hence

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}}\|\boldsymbol{y}_{\text{LASER}}(\boldsymbol{M}_r) - \boldsymbol{M}_r\|^2 &= \mathbb{E}_{\boldsymbol{Z}}\|\sum_i \widetilde{\boldsymbol{p}}_i \widetilde{\boldsymbol{q}}_i^\top\|^2 - \|\boldsymbol{M}_r\|^2 \\
&= \sum_i \mathbb{E}_{\boldsymbol{Z}}\|\widetilde{\boldsymbol{p}}_i\|^2 \, \mathbb{E}_{\boldsymbol{Z}}\|\widetilde{\boldsymbol{q}}_i\|^2 - \sum_i \|\boldsymbol{p}\|^2\|\boldsymbol{q}\|^2 \\
&= \sum_i \|\boldsymbol{p}\|^2\|\boldsymbol{q}\|^2 \left[\left(1 + \frac{m}{\alpha_i}\right)\left(1 + \frac{n}{\beta_i}\right) - 1\right] \\
&= \|\boldsymbol{M}_r\|^2 \left(\sum_i \kappa_i \left(1 + \frac{m}{\alpha_i}\right)\left(1 + \frac{n}{\beta_i}\right) - 1\right) \\
&\overset{(Lemma\ 8)}{=} \|\boldsymbol{M}_r\|^2 \left(f_P(\boldsymbol{\alpha}, \boldsymbol{\beta}) - 1\right),
\end{aligned}$$

where we set $\kappa_i = \|\boldsymbol{p}_i\|^2 / \|\boldsymbol{M}_r\|^2$. Now choosing $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ as in Lemma 8 yields the desired result. □

E.5 Proof of Lemma 11

Proof. From Eq. (14), we have that

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \gamma_t \boldsymbol{g}_t + \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)\right).$$

Denoting $\text{Error}_{\boldsymbol{Z}} = \mathcal{C}_\text{r}(\boldsymbol{e}_\text{t} + \gamma_\text{t} \boldsymbol{g}_\text{t}) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_\text{r}(\boldsymbol{e}_\text{t} + \gamma_\text{t} \boldsymbol{g}_\text{t})\right)$, we observe that $\mathbb{E}_{\boldsymbol{Z}}[\text{Error}_{\boldsymbol{Z}}] = 0$ and $\mathbb{E}_{\boldsymbol{Z}}\|\text{Error}_{\boldsymbol{Z}}\|^2 \leq \lambda_{\text{LASER}}\|\mathcal{C}_\text{r}(\boldsymbol{e}_\text{t} + \gamma_\text{t} \boldsymbol{g}_\text{t})\|^2 \leq \lambda_{\text{LASER}}\|\boldsymbol{e}_\text{t} + \gamma_\text{t} \boldsymbol{g}_\text{t}\|^2$ (see App. E.4). Thus

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \\
&= \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star - \gamma_t \boldsymbol{g}_t\|^2 + \mathbb{E}\|\text{Error}_{\boldsymbol{Z}}\|^2 \\
&= \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - 2\gamma_t \mathbb{E}\langle \boldsymbol{g}_t, \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\rangle + \gamma_t^2 \mathbb{E}\|\boldsymbol{g}_t\|^2 + \mathbb{E}\|\text{Error}_{\boldsymbol{Z}}\|^2 \\
&\leq \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - 2\gamma_t \mathbb{E}\langle \boldsymbol{g}_t, \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\rangle + 2\gamma_t \mathbb{E}\langle \boldsymbol{g}_t, \boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\rangle + \gamma_t^2 \mathbb{E}\|\boldsymbol{g}_t\|^2 + \lambda_{\text{LASER}} \mathbb{E}\|\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t\|^2 \\
&= \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - 2\gamma_t \mathbb{E}\langle \boldsymbol{g}_t, \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\rangle + 2\gamma_t \mathbb{E}\langle \boldsymbol{g}_t, \boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\rangle(1 + \lambda_{\text{LASER}}) + \gamma_t^2 \mathbb{E}\|\boldsymbol{g}_t\|^2(1 + \lambda_{\text{LASER}}) \\
&\qquad + \lambda_{\text{LASER}} \mathbb{E}\|\boldsymbol{e}_t\|^2 \\
&\overset{(Assump.\ 3)}{\leq} \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - 2\gamma_t \mathbb{E}\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\rangle + 2\gamma_t \mathbb{E}\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\rangle(1 + \lambda_{\text{LASER}}) \\
&\qquad + (M+1)(1 + \lambda_{\text{LASER}})\gamma_t^2 \mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \gamma_t^2 \sigma^2 (1 + \lambda_{\text{LASER}}) + \lambda_{\text{LASER}} \mathbb{E}\|\boldsymbol{e}_t\|^2. \qquad (21)
\end{aligned}$$

Now we closely follow the steps as in the proof of [48, Lemma 8]. Since f is L-smooth, we have $\|\nabla f(\boldsymbol{\theta}_t)\|^2 \leq 2L(f(\boldsymbol{\theta}_t) - f_\star$. Further, by Assumption 1,

$$-2\langle \nabla f(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\rangle \leq -\mu\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\|^2 - 2(f(\boldsymbol{\theta}_t) - f_\star),$$

and since $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \alpha\|\boldsymbol{a}\|^2 + \alpha^{-1}\|\boldsymbol{b}\|^2$ for $\alpha > 0$, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$, we have

$$2\langle \nabla f(\boldsymbol{\theta}_t), \widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\rangle \leq \frac{1}{2L}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + 2L\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 \leq f(\boldsymbol{\theta}_t) - f_\star + 2L\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 .$$

And by $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq (1+\beta)\|\boldsymbol{a}\|^2 + (1+\beta^{-1})\|\boldsymbol{b}\|^2$ for $\beta > 0$ (via Jensen's inequality), we observe

$$-\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star\|^2 \leq -\frac{1}{2}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 + \|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 .$$

Plugging these inequalities in Eq. (21), we obtain that

$$\begin{aligned}
&\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \\
&\leq \left(1 - \frac{\mu\gamma_t}{2}\right)\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - \gamma_t\left(1 - \lambda_{\text{LASER}} - 2L(M+1)(1+\lambda_{\text{LASER}})\gamma_t\right)\mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star) \\
&\qquad + \gamma_t^2\sigma^2(1+\lambda_{\text{LASER}}) + (\mu\gamma_t + 2L\gamma_t(1+\lambda_{\text{LASER}}))\mathbb{E}\|\boldsymbol{e}_t\|^2.
\end{aligned}$$

Utilizing the fact that $\gamma_t \leq \frac{1-2\lambda_{\text{LASER}}}{4L(M+1)(1+\lambda_{\text{LASER}})}$ and $\mu \leq L$ yields the desired claim. □

E.6 Proof of Lemma 12

Proof. The proof of Lemma 12 is very similar to that of Lemma 2 for EF-SGD. In that proof, a key step is to establish that $(3L(2/\delta + M)\gamma_t^2) \leq \frac{\delta}{64L}$ and $(3L\gamma_t\, 4/\delta) \leq 1$. In our setting, $\gamma_t \leq \frac{1}{10L(2/\delta_r + M)(1+\lambda_{\text{LASER}})}$ and $\lambda_{\text{LASER}} \leq \frac{1}{10(2/\delta_r + M)}$. Thus

$$\begin{aligned}
&\left(3L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\gamma_t^2\left(\frac{2}{\delta_r} + M\right) \\
&= 3L\left(\frac{2}{\delta_r} + M\right)(1+\lambda_{\text{LASER}})\gamma_t \cdot \gamma_t + \lambda_{\text{LASER}}\left(\frac{2}{\delta_r} + M\right)\gamma_t \\
&\leq \frac{3}{10}\cdot\gamma_t + \frac{1}{10}\cdot\gamma_t \\
&= \frac{4}{10}\,\frac{1}{10L(\frac{2}{\delta_r} + M)(1+\lambda_{\text{LASER}})} \\
&\leq \frac{\delta_r}{32L}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{4}{\delta_r}\left(3L(1+\lambda_{\text{LASER}})\gamma_t + \lambda_{\text{LASER}}\right) &= 3L(1+\lambda_{\text{LASER}})\,\frac{4}{\delta_r}\,\gamma_t + \lambda_{\text{LASER}}\,\frac{4}{\delta_r} \\
&\leq \frac{6}{10} + \frac{2}{10} \\
&\leq 1.
\end{aligned}$$

□

E.7 Proof of Lemma 13

Proof. From Eq. (14), we have that

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \gamma_t\boldsymbol{g}_t + \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t\boldsymbol{g}_t) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t\boldsymbol{g}_t)\right).$$

Denoting $\text{Error}_{\boldsymbol{Z}} = \mathcal{C}_{\text{r}}(\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}}) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_{\text{r}}(\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}})\right)$, we observe that $\mathbb{E}_{\boldsymbol{Z}}[\text{Error}_{\boldsymbol{Z}}] = 0$ and $\mathbb{E}_{\boldsymbol{Z}}\|\text{Error}_{\boldsymbol{Z}}\|^2 \leq \lambda_{\text{LASER}}\|\mathcal{C}_{\text{r}}(\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}})\|^2 \leq \lambda_{\text{LASER}}\|\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}}\|^2$ (see App. E.4). Using the smoothness of f,

$$f(\widetilde{\boldsymbol{\theta}}_{t+1}) \leq f(\widetilde{\boldsymbol{\theta}}_t) - \gamma_t\langle\nabla f(\widetilde{\boldsymbol{\theta}}_t), \boldsymbol{g}_t\rangle + \langle f(\widetilde{\boldsymbol{\theta}}_t), \text{Error}_{\boldsymbol{Z}}\rangle + \frac{L}{2}\| - \gamma_t\boldsymbol{g}_t + \text{Error}_{\boldsymbol{Z}}\|^2$$

Taking expectation on both sides,

$$\mathbb{E}f(\widetilde{\boldsymbol{\theta}}_{t+1}) \leq \mathbb{E}f(\widetilde{\boldsymbol{\theta}}_t) - \gamma_t\mathbb{E}\langle\nabla f(\widetilde{\boldsymbol{\theta}}_t), \nabla f(\boldsymbol{\theta}_t)\rangle + \frac{L}{2}\left(\gamma_t^2\mathbb{E}\|\boldsymbol{g}_t\|^2 + \lambda_{\text{LASER}}\mathbb{E}\|\boldsymbol{e}_t + \gamma_t\boldsymbol{g}_t\|^2\right).$$

Rewriting $\langle \nabla f(\widetilde{\boldsymbol{\theta}}_t), \nabla f(\boldsymbol{\theta}_t)\rangle = \|\nabla f(\boldsymbol{\theta}_t)\|^2 + \langle \nabla f(\widetilde{\boldsymbol{\theta}}_t) - \nabla f(\boldsymbol{\theta}_t), \nabla f(\boldsymbol{\theta}_t)\rangle$ and using $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$, we can simplify the expression as

$$\begin{aligned}\langle \nabla f(\widetilde{\boldsymbol{\theta}}_t) - \nabla f(\boldsymbol{\theta}_t), \nabla f(\boldsymbol{\theta}_t)\rangle &\leq \frac{1}{2}\|\nabla f(\boldsymbol{\theta}_t) - \nabla f(\widetilde{\boldsymbol{\theta}}_t)\|^2 + \frac{1}{2}\|\nabla f(\boldsymbol{\theta}_t)\|^2 \\ &\leq \frac{L^2}{2}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 + \frac{1}{2}\|\nabla f(\boldsymbol{\theta}_t)\|^2 .\end{aligned}$$

Pluggin this inequality back together with $\mathbb{E}\|\boldsymbol{g}_t\|^2 \leq (M+1)\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \sigma^2$, we get

$$\begin{aligned}\mathbb{E}f(\widetilde{\boldsymbol{\theta}}_{t+1}) \leq \mathbb{E}f(\widetilde{\boldsymbol{\theta}}_t) - \frac{\gamma_t}{2}\left(1 - 2\gamma_t L(M+1)(1+\lambda_{\text{LASER}})\right)\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{L\gamma_t^2\sigma^2(1+\lambda_{\text{LASER}})}{2} \\ + L\left(\frac{L\gamma_t}{2} + \lambda_{\text{LASER}}\right)\mathbb{E}\|\boldsymbol{e}_t\|^2.\end{aligned}$$

Now utilizing the fact $\gamma_t \leq \frac{1}{4L(M+1)(1+\lambda_{\text{LASER}})}$ establishes the desired result. □

F Additional details about noisy channel and LASER

F.1 Channel transformation

Recall from Eq. (2) in Sec. 2 that the server first obtains $\boldsymbol{y} = \sum_{i=1}^{k} a_i \boldsymbol{g}_i + \boldsymbol{Z}$, where $\|a_i \boldsymbol{g}_i\|^2 \leq P$ (note that we use the constant scheme $P_t = P$ as justified in Sec. 4.2). Now we want to show that for estimating the gradient sum $\sum_i \boldsymbol{g}_i$ through a linear transformation on $\boldsymbol{y}$, the optimal power scalars are given by $a_i = \frac{\sqrt{P}}{\max_j \|\boldsymbol{g}_j\|}$, $\forall i \in [k]$, which yields the channel model in (noisy channel).

Towards this, first let $k = 2$ (the proof for general k is similar). Thus our objective is

$$\min_{a_1, a_2, b} \mathbb{E}\left\|\frac{\boldsymbol{y}}{b} - \boldsymbol{g}_1 - \boldsymbol{g}_2\right\|^2 .$$

For any a_1, a_2, b, we have that

$$\begin{aligned}\mathbb{E}\left\|\frac{\boldsymbol{y}}{b} - \boldsymbol{g}_1 - \boldsymbol{g}_2\right\|^2 &= \min_{a_1,a_2,b:\|a_i\boldsymbol{g}_i\|^2\leq P} \mathbb{E}\left\|\boldsymbol{g}_1\left(\frac{a_1}{b} - 1\right) + \boldsymbol{g}_2\left(\frac{a_2}{b} - 1\right) + \frac{\boldsymbol{Z}}{b}\right\|^2 \\ &= \min_{a_1,a_2,b:\|a_i\boldsymbol{g}_i\|^2\leq P} \mathbb{E}\left\|\nabla f(\boldsymbol{\theta})(\Delta_1 + \Delta_2) + \Delta_1 \boldsymbol{\xi}_1 + \Delta_2 \boldsymbol{\xi}_2 + \frac{\boldsymbol{Z}}{b}\right\|^2, \quad \Delta_i = \frac{a_i}{b} - 1 \\ &= \min_{a_1,a_2,b:\|a_i\boldsymbol{g}_i\|^2\leq P} \left(\|\nabla f(\boldsymbol{\theta})\|^2(\Delta_1+\Delta_2)^2 + \Delta_1^2\,\mathbb{E}\|\boldsymbol{\xi}_1\|^2 + \Delta_2^2\,\mathbb{E}\|\boldsymbol{\xi}_2\|^2 + \frac{\mathbb{E}\|\boldsymbol{Z}\|^2}{b^2}\right),\end{aligned}$$

where we used the fact that $\boldsymbol{g}_1 = \nabla f(\boldsymbol{\theta}) + \boldsymbol{\xi}_1$ and $\boldsymbol{g}_2 = \nabla f(\boldsymbol{\theta}) + \boldsymbol{\xi}_2$ with zero-mean and independent $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2$, and $\boldsymbol{Z}$. We now observe that for any fixed b the optimal a_i's are given by $a_1 = a_2 = b$, i.e. $\Delta_1 = \Delta_2 = 0$. To determine the optimal b, we have to solve

$$\max\ b \quad \text{s.t.}\ \|b\,\boldsymbol{g}_i\|^2 \leq P,$$

which yields $b^\star = \sqrt{P}/\max_i \|\boldsymbol{g}_i\|$. The proof for general k is similar.

F.2 Detailed steps for Algorithm 1

Now we delineate the two main components of LASER: (i) Gradient compression + Error-feedback (EF), and (ii) Power allocation + Channel transmission.

Gradient compression and error feedback (7-9). Since we transmit low-rank gradient approximations, we use error feedback (EF) to incorporate the previous errors into the current gradient update. This ensures convergence of SGD with biased compressed gradients [49]. For the rank-r compression of the updated gradient $\boldsymbol{M}$, $\mathcal{C}_r(\boldsymbol{M})$, we use the PowerSGD algorithm from [12], a linear compression scheme to compute the left and right singular components $\boldsymbol{P} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times r}$ respectively. PowerSGD uses a single step of the subspace iteration [50] with a warm start from the previous updates to compute these factors. The approximation error, $\boldsymbol{M} - \boldsymbol{P}\boldsymbol{Q}^\top$, is then used to update the error-feedback for next iteration. Note that the clients do not have access to the channel output and only include the local compression errors into their feedback. The decompression function in line 9 is given by $\text{DECOMPRESS}(\boldsymbol{P}, \boldsymbol{Q}) \triangleq \boldsymbol{P}\boldsymbol{Q}^\top \in \mathbb{R}^{m \times n}$.

Power allocation and channel transmission (10-11). For each client, to transmit the rank-r factors $\boldsymbol{P}$ and $\boldsymbol{Q}$ over the noisy channel, we compute the corresponding power-allocation vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}_+^r$, given by $\boldsymbol{\alpha}, \boldsymbol{\beta} = \text{POWERALLOC}(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{M})$. This allocation is uniform across all the clients. Given these power scalars, all the clients synchronously transmit the corresponding left factors over the channel which results in $\boldsymbol{Y}_{\boldsymbol{p}} \in \mathbb{R}^{m \times r}$. Similarly for $\boldsymbol{Y}_{\boldsymbol{q}} \in \mathbb{R}^{n \times r}$. Finally, the stochastic gradient for the model update is reconstructed as $\boldsymbol{g} = \boldsymbol{Y}_{\boldsymbol{p}} \boldsymbol{Y}_{\boldsymbol{q}}^\top$. For brevity we defer the full details to § F.1.

Recall from Algorithm 1 that power allocation among clients is done via the function $\text{POWERALLOC}(\{\mathcal{C}_r(\boldsymbol{M}_j), \boldsymbol{M}_j\})$. The theoretically optimal power allocation is discussed in App. C.1, and given explicitly in Lemma 8. However we empirically observe that we can relax this allocation scheme and even simpler schemes suffice to beat the other considered baselines. This is detailed in App. G.5.

F.3 Constant-order SNR

As discussed in Sec. 3.2 and established in Lemmas 9 and 10 of App. C.2, we have that

$$\lambda_{\text{LASER}} \leq \frac{4}{(m/r)\,\text{SNR}}\left(1+\frac{1}{(n/r)\,\text{SNR}}\right) \ll \frac{1}{\text{SNR}} = \lambda_{\text{Z-SGD}}.$$

In the low-rank [12] and constant-order SNR regime where $r = \mathcal{O}(1)$ and $\text{SNR} = \Omega(1)$, we observe that λ_{LASER} is roughly $\mathcal{O}(m)$ times smaller than $\lambda_{\text{Z-SGD}}$. In other words, the effective SNR seen by LASER roughly gets boosted to $\mathcal{O}(m\,\text{SNR})$ due to capitalizing on the low-rank factors whereas Z-SGD perceives only the standard factor SNR. Constant-order SNR, i.e. $P/mn = \Omega(1)$, means that the energy used to transmit each coordinate is roughly a constant, analogous to the constant-order bits used in quantization schemes [37]. Note that this is only a sufficient theoretical condition to ensure that the ratio between λ_{LASER} and $\lambda_{\text{Z-SGD}}$ is smaller than one. In fact, a much weaker condition that $P/4r^2 > 1$ suffices. To establish this, we note

$$\frac{\lambda_{\text{LASER}}}{\lambda_{\text{Z-SGD}}} = \frac{4r}{m}\left(1+\frac{r}{n\text{SNR}}\right) = \frac{4r}{m}\left(1+\frac{rm}{P}\right) = \frac{4r}{m} + \frac{4r^2}{P}.$$

The first term is usually negligible since we always fix the rank $r = 4$, which is much smaller compared to m in the architectures we consider. Thus if $P/4r^2 > 1$, we see that the above ratio is smaller than one. Note that the constant-order SNR assumption already guarantees this: $\text{SNR} = \Omega(1) \Rightarrow P \gtrsim mn \Rightarrow P \gtrsim r^2$, since r is smaller than both m and n. On the other hand, for the RESNET18 architecture with $L = 61$ layers and $r = 4$, the power levels $P = 250, 500$ violate the above condition as $P/(Lr^2) < 4$ (note that the budget P here is for the entire network and hence replaced by P/L). But empirically we still observe the accuracy gains in this low-power regime (Fig. 2 in the paper).

G Experimental details

We provide technical details for the experiments demonstrated in Sec. 4. First we provide the general description followed by each individual experimental setup.

General setup. We consider four challenging tasks of practical interest: (i) GPT language modeling on WIKITEXT-103, and (ii, iii, iv) image classification on MNIST, CIFAR10 and CIFAR100. For the language modeling, we use the GPT-2 like architecture following [51] (§ G). RESNET18 is used for the CIFAR datasets. For MNIST, we use a 1-hidden-layer network for a fair comparison with [15]. For distributed training of these models, we consider $k = 4$ clients for language modeling and $k = 16$ for image classification. We simulate the noisy channel by sampling $\boldsymbol{Z} \sim \mathcal{N}(0, \mathbf{I}_d)$. To gauge the performance of algorithms over a wide range of noisy conditions, we vary the power P geometrically in the range $[0.1, 10]$ for MNIST, $[250, 128000]$ for CIFAR10 and CIFAR100, and $[10000, 1024 \times 10000]$ for WIKITEXT-103. The chosen ranges can be roughly split into low-moderate-high power regimes. Recall from noisy channel that the smaller the power, the higher the noise in the channel.

Baselines. We benchmark LASER against three different sets of baselines: (i) Z-SGD, (ii) SIGNUM, RANDOM-K, SKETCHING, and (iii) A-DSGD. Z-SGD sends the uncompressed gradients directly over the noisy channel and acts as a canonical baseline. The algorithms in (ii) are state-of-the-art distributed compression schemes for noiseless communication [12]. SIGNUM [5] transmits the gradient sign followed by the majority vote and SKETCHING [52, 53] uses a Count Mean Sketch to compress the gradients. We omit comparison with quantization methods [6] given the difference in our objectives and the settings (noisy channel). A-DSGD [15] is a popular compression scheme for noisy channels, relying on Top-K and random sketching. However A-DSGD does not scale to tasks of the size we consider and hence we benchmark against it only on MNIST. SGD serves as the noiseless baseline (Table 2). All the compression algorithms use the error-feedback, and use the compression factor (compressed-gradient-size/original-size) 0.2, the optimal in the range $[0.1, 0.8]$. We report the best results among 3 independent runs for all the baselines.

G.1 WIKITEXT-103 experimental setup

This section concerns the experimental details used to obtain Fig. 1 and Table 1 in the main text. Table 4 collects the settings we adopted to run our code. Table 5 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 4: Default experimental settings for the GPT-2 model used to learn the WIKITEXT-103 task.

Dataset	WIKITEXT-103
Architecture	GPT-2 (as implemented in [51])
Number of workers	4
Batch size	15 per worker
Accumulation steps	3
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)
Learning rate	0.001
Scheduler	Cosine
# Iterations	20000
Weight decay	1×10^{-3}
Dropout	0.2
Sequence length	512
Embeddings	768
Transformer layers	12
Attention heads	12
Power budget	6 levels: 10k, 40k, 160k, 640k, 2560k, 10240k
Power allocation	Proportional to norm of compressed gradients (uncompressed gradients for Z-SGD)
Compression	Rank 4 for LASER; 0.2 compression factor for other baselines
Repetitions	1

Table 5: Parameters in the GPT-2 architecture, with their shape and uncompressed size.

Parameter	Gradient tensor shape	Matrix shape	Uncompressed size
transformer.wte	50304×768	50304×768	155 MB
transformer.wpe	512×768	512×768	1573 KB
transformer.h.ln_1 ($\times 12$)	768	768×1	($12\times$) 3 KB
transformer.h.attn.c_attn ($\times 12$)	2304×768	2304×768	($12\times$) 7078 KB
transformer.h.attn.c_proj ($\times 12$)	768×768	768×768	($12\times$) 2359 KB
transformer.h.ln_2 ($\times 12$)	768	768×1	($12\times$) 3 KB
transformer.h.mlp.c_fc ($\times 12$)	3072×768	3072×768	($12\times$) 9437 KB
transformer.h.mlp.c_proj ($\times 12$)	768×3072	768×3072	($12\times$) 9437 KB
transformer.ln_f	768	768×1	3 KB
Total			496 MB

G.2 CIFAR10 experimental results

This section concerns additional results on CIFAR10, as well as the experimental details used to obtain Fig. 2. Table 6 highlights that LASER requires just $(1/16)^{\text{th}}$ the power compared to Z-SGD to reach any target accuracy till 91%. Table 7 collects the settings we adopted to run our code. Table 8 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 6: Power required (*lower the better*) to reach the given target accuracy on CIFAR-10. LASER requires $16\times$ lesser power than the Z-SGD to achieve the same targetaccuracy. Equivalently, LASER tolerates more channel noise than the Z-SGD for the same target accuracy as is partly supported by our theoretical analysis.

Target	**Power required**		**Reduction**
	LASER	Z-SGD	
88%	250	4000	$16\times$
89%	500	8000	$16\times$
90%	1000	16000	$16\times$
91%	2000	32000	$16\times$

Table 7: Default experimental settings for the RESNET18 model used to learn the CIFAR10 task.

Dataset	CIFAR10
Architecture	RESNET18
Number of workers	16
Batch size	128 per worker
Optimizer	SGD
Momentum	0.9
Learning rate	Grid-searched in $\{0.001, 0.005, 0.01, 0.05\}$ for each power level
# Epochs	150
Weight decay	1×10^{-4}, 0 for BatchNorm parameters
Power budget	10 levels: $250, 500, 1000, 2000, 4000, 8000, 16000, 32000, 64000, 128000$
Power allocation	Proportional to norm of compressed gradients (uncompressed gradients for Z-SGD)
Compression	Rank 4 for LASER; 0.2 compression factor for other baselines
Repetitions	3, with varying seeds

Table 8: Parameters in the ResNet18 architecture, with their shape and uncompressed size.

Parameter	Gradient tensor shape	Matrix shape	Uncompressed size
layer4.1.conv2	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.0.conv2	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.1.conv1	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.0.conv1	$512 \times 256 \times 3 \times 3$	512×2304	4719 KB
layer3.1.conv2	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.1.conv1	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.0.conv2	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.0.conv1	$256 \times 128 \times 3 \times 3$	256×1152	1180 KB
layer2.1.conv2	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer2.1.conv1	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer2.0.conv2	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer4.0.shortcut.0	$512 \times 256 \times 1 \times 1$	512×256	524 KB
layer2.0.conv1	$128 \times 64 \times 3 \times 3$	128×576	295 KB
layer1.1.conv1	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.1.conv2	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.0.conv2	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.0.conv1	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer3.0.shortcut.0	$256 \times 128 \times 1 \times 1$	256×128	131 KB
layer2.0.shortcut.0	$128 \times 64 \times 1 \times 1$	128×64	33 KB
linear	10×512	10×512	20 KB
conv1	$64 \times 3 \times 3 \times 3$	64×27	7 KB
Bias vectors (total)			38 KB
Total			45 MB

G.3 CIFAR100 experimental results

This section concerns experimental results on CIFAR100. We used the same RESNET18 architecture as for CIFAR10 (except for the final layer, adapted to the 100-class dataset). We once again compared LASER to the usual baselines. Fig. 4 and Table 11 collect the results that we obtained. It can be seen that LASER outperforms the other algorithms with an even wider margin compared to the CIFAR10 and WIKITEXT-103 tasks, with a power gain of around $32\times$ across different accuracy targets. SIGNUM is much more sensitive to noise and performs much worse than the other algorithms; therefore, we decided to leave out its results in order to improve the quality of the plot. Table 9 collects the settings we adopted to run our code. Table 10 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 9: Default experimental settings for the RESNET18 model used to learn the CIFAR100 task.

Dataset	CIFAR100
Architecture	RESNET18
Number of workers	16
Batch size	128 per worker
Optimizer	SGD
Momentum	0.9
Learning rate	Grid-searched in $\{0.001, 0.005, 0.01, 0.05\}$ for each power level
LR decay	$/10$ at epoch 150
# Epochs	200
Weight decay	1×10^{-4}
	0 for BatchNorm parameters
Power budget	10 levels: $500, 1000, 2000, 4000, 8000, 16000, 32000, 64000, 128000, 256000$
Power allocation	Proportional to norm of compressed gradients (uncompressed gradients for Z-SGD)
Repetitions	3, with varying seeds
Compression	Rank 4 for LASER; 0.2 compression factor for other baselines

Table 10: Parameters in the ResNet18 architecture, with their shape and uncompressed size.

Parameter	Gradient tensor shape	Matrix shape	Uncompressed size
layer4.1.conv2	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.0.conv2	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.1.conv1	$512 \times 512 \times 3 \times 3$	512×4608	9437 KB
layer4.0.conv1	$512 \times 256 \times 3 \times 3$	512×2304	4719 KB
layer3.1.conv2	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.1.conv1	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.0.conv2	$256 \times 256 \times 3 \times 3$	256×2304	2359 KB
layer3.0.conv1	$256 \times 128 \times 3 \times 3$	256×1152	1180 KB
layer2.1.conv2	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer2.1.conv1	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer2.0.conv2	$128 \times 128 \times 3 \times 3$	128×1152	590 KB
layer4.0.shortcut.0	$512 \times 256 \times 1 \times 1$	512×256	524 KB
layer2.0.conv1	$128 \times 64 \times 3 \times 3$	128×576	295 KB
layer1.1.conv1	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.1.conv2	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.0.conv2	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer1.0.conv1	$64 \times 64 \times 3 \times 3$	64×576	147 KB
layer3.0.shortcut.0	$256 \times 128 \times 1 \times 1$	256×128	131 KB
layer2.0.shortcut.0	$128 \times 64 \times 1 \times 1$	128×64	33 KB
linear	100×512	100×512	205 KB
conv1	$64 \times 3 \times 3 \times 3$	64×27	7 KB
Bias vectors (total)			38 KB
Total			45 MB

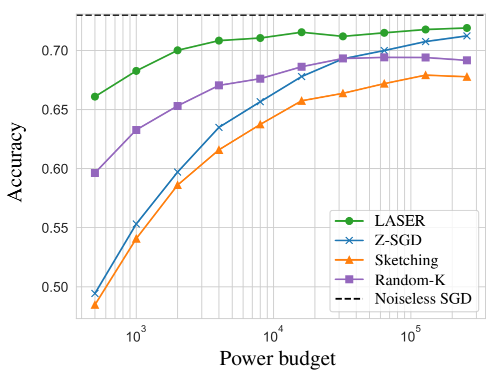

Figure 4: Test accuracy (*higher the better*) for a given power budget on CIFAR-100 for different algorithms. The advantage of LASER is evident across the entire power spectrum.

Table 11: Power required (*lower the better*) to reach the given target accuracy on CIFAR-100. LASER requires $16 - 32\times$ lesser power than the Z-SGD to achieve the same targetaccuracy. Equivalently, LASER tolerates more channel noise than the Z-SGD for the same target accuracy as is partly supported by our theoretical analysis.

Target	**Power required**		**Reduction**
	LASER	Z-SGD	
65%	500	8000	$16\times$
68%	1000	32000	$32\times$
70%	2000	64000	$32\times$
71%	8000	256000	$32\times$

G.4 MNIST experimental results

Table 12 collects the settings we adopted to run our code.

Table 12: Default experimental settings for the 1-LAYER NN used to learn the MNIST task.

Dataset	MNIST
Architecture	1-LAYER NN
Number of workers	16
Batch size	128 per worker
Optimizer	SGD
Momentum	0.9
Learning rate	0.01
# Epochs	50
Weight decay	1×10^{-4},
Power budget	3 levels: $0.1, 1, 10$
Power allocation	Proportional to norm of compressed gradients (uncompressed gradients for Z-SGD)
Repetitions	3, with varying seeds
Compression	Rank 2 for LASER; 0.1 compression factor for other baselines

Table 13: Test accuracy (*higher the better*) after 50 epochs on MNIST for low, moderate, and high power regimes.

Algorithm	Test accuracy		
	$P = 0.1$	$P = 1$	$P = 10$
Z-SGD	81.3%	87.9%	91.9%
SIGNUM	76.7%	83.2%	85.4%
RANDOM-K	**86.1**%	89.3%	91.5%
SKETCHING	81.9%	88.2%	91.7%
A-DSGD	81.6%	86.9%	87.3%
LASER	84.3%	**89.9**%	**92.3**%

Table 13 compares the performance of LASER against various compression algorithms on MNIST. In the very noisy regime ($P = 0.1$), RANDOM-K is slightly better than LASER and outperforms the other baselines, whereas in the moderate ($P = 1$) and high power ($P = 10$) regimes, LASER is slightly better than the other algorithms. On the other hand, we observe that A-DSGD performs worse than even simple compression schemes like RANDOM-K in all the settings.

G.5 Power allocation across workers and neural network parameters

The choice of power allocation over the layers of the network is perhaps the most important optimization required in our experimental setup. Notice that, because of Eq. (2), all clients must allocate the same power to a given gradient, since otherwise it would be impossible to recover the correct average gradient. However, workers have a degree of freedom in choosing how to distribute the power budget among gradients, i.e. among the layers of the network, and this power allocation can change over the iterations of the model training.

App. C.1 analyzes power allocation optimality from a theoretical point of view. On the experimental side, simpler schemes are enough to get significant gains over the other baselines. As a matter of fact, we considered the following power allocation scheme for the experiments: at each iteration, each worker determines locally how to allocate its power budget across the gradients. Then, we assume that this power allocation choice is communicated by the client to the server noiselessly. The server then takes the average of the power allocation choices, and communicates the final power allocation to the clients. The clients then use this power allocation to send the gradients to the server via the noisy channel.

For the determination of each worker's power allocation, three schemes were considered:

- uniform power to each gradient;
- power proportional to the Frobenius norm (or the square of it) of the gradients;
- power proportional to the norm of the compressed gradients (i.e., the norm of what is actually communicated to the server).

For Z-SGD, where there is no gradient compression, the best power allocation turned out to be the one proportional to the norm of the gradients, independently of the power constraint imposed. For all the other algorithms, the best is power proportional to the norm of the compressed gradients.

G.6 Baselines implementation

In this section we describe our implementation of the baselines considered in the paper.

G.6.1 Count-Mean Sketching

Algorithm 2 COUNT-MEAN SKETCHING

function COMPRESS(gradient matrix $M \in \mathbb{R}^{n \times m}$)
  Treat M as a vector of length nm.
  The number of samples b is set to $mn \times$ (compression factor).
  If the resulting b is less than 1, we set $b = 1$.
  Sample a set of mn indices I i.i.d. between 0 and $b - 1$ using the same seed on all workers.
  Sample a set of mn signs ($+1$ or -1) S i.i.d. using the same seed used for I.
  $\hat{C} \leftarrow \mathbf{0} \in \mathbb{R}^b$
  for $j = 0, \ldots, mn - 1$ **do**
    $\hat{C}(I(j)) \leftarrow \hat{C}(I(j)) + S(j) \times M(j)$
  end for
  return $\hat{C}$
end function
function AGGREGATE+DECOMPRESS(worker's values $\hat{C}_1 \ldots \hat{C}_k$)
  Sample I and S as before, using the same seed.
  $\hat{M} \leftarrow \mathbf{0} \in \mathbb{R}^{n \times m}$
  $\hat{M}(I) \leftarrow \frac{1}{k} \sum_{i=1}^{k} \hat{C}_i(I) \odot S$
  return $\hat{M}$
end function

Power is allocated proportional to compressed gradients' norms. The algorithm is implemented without local error feedback, since error feedback causes the algorithm to diverge. The compression factor was grid-searched in $\{0.1, 0.2, 0.5, 0.8\}$ and 0.2 was finally chosen as the overall best.

G.6.2 Random K

Algorithm 3 Random K

function COMPRESS(gradient matrix $M \in \mathbb{R}^{n \times m}$)
  Treat M as a vector of length nm.
  The number of samples b is set to $mn \times$ (compression factor).
  If the resulting b is less than 1, we set $b = 1$.
  Sample a set of b indices I without replacement, using the same seed on all workers.
  return Looked up values $S = M(I)$.
end function
function AGGREGATE+DECOMPRESS(worker's values $S_1 \dots S_k$)
  $\hat{M} \leftarrow \mathbf{0} \in \mathbb{R}^{n \times m}$
  $\hat{M}(I) \leftarrow \frac{1}{k} \sum_{i=1}^{k} S_i$
  return $\hat{M}$
end function

Power is allocated proportional to compressed gradients' norms. The algorithm is implemented with local error feedback. The compression factor was grid-searched in $\{0.1, 0.2, 0.5, 0.8\}$ and 0.2 was finally chosen as the overall best.

G.6.3 Signum

Algorithm 4 SIGNUM

function COMPRESS(gradient matrix $M \in \mathbb{R}^{n \times m}$)
  Compute the signs $S \in \{-1, 1\}^{n \times m}$ of M
  return S
end function
function AGGREGATE+DECOMPRESS(worker's signs $S_1 \dots S_k$)
  return SIGN($\sum_{i=1}^{k} S_i$)
end function

We implemented SIGNUM following [5]. We run it in its original form, without error feedback. Power is allocated proporional to the compressed gradients' norms. Since the compressed gradients are simply the sign matrices, in this case power is allocated proportional to the square root of the number of parameters in each layer $\sqrt{mn}$. Unlike the other baselines, SIGNUM does not require any compression factor.

H Analysis

H.1 Rank-accuracy tradeoff

There exists an inherent tradeoff between the decomposition rank r (and hence the compression factor δ_r) and the final model accuracy. In fact, a small rank r implies aggressive compression and hence the compression noise dominates the channel noise. Similarly, for a high decomposition rank, the channel noise overpowers the compression noise as the power available per each coordinate is small. We empirically investigate this phenomenon for CIFAR10 classification over various power regimes in Fig. 5.

As Fig. 5 reveals, either Rank-4 or Rank-8 compression is optimal for all the three power regimes. Further we observe two interesting trends: (i) the final accuracy is uniformly worse in all the regimes with overly aggressive rank-one compression, and (ii) higher rank compression impacts the low power regime more significantly than the medium and high-power counterparts. This is in agreement with the intuition that at low power (and hence noisier channel), it is better to allocate the limited power budget appropriately to few "essential" rank components as opposed to thinning it out over many. This phenomenon can be theoretically explained by characterizing the compression factor δ_r

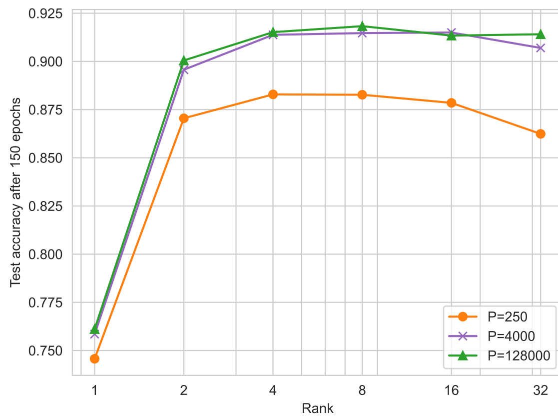

Figure 5: Final accuracy vs. compression rank tradeoff for CIFAR-10 classification, for low, medium and high power regimes. Rank-4/Rank-8 compression is optimal for all the three regimes. It reveals two interesting insights: (i) performance is uniformly worse in all the regimes with overly aggressive rank-one compression, and (ii) higher rank compression impacts low power regime more significantly than the medium and high-power counterparts. This confirms with the intuition that at low power (and hence noisier channel), it is better to allocate the limited power budget appropriately to few "essential" rank components as opposed to thinning it out over many.

as a function of rank r and its effect on the model convergence. While the precise expression for δ_r is technically challenging, given the inherent difficulty in analyzing the PowerSGD algorithm [12], we believe that a tractable characterization of this quantity (via upper bounds etc.) can offer fruitful insights into the fundamental rank-accuracy tradeoff at play.

To further shed light on this phenomenon, we trained the noiseless SGD on CIFAR10 and captured the evolution across the epochs of the energy contained in the top eight components of each gradient matrix. As illustrated in Fig. 6, we observe that for the first and last hidden layers, 80% of the energy is already captured in these eight components. On the other hand, for the middle layer this fares around 55%. It is interesting to further explore this behavior for GPT models and other tasks.

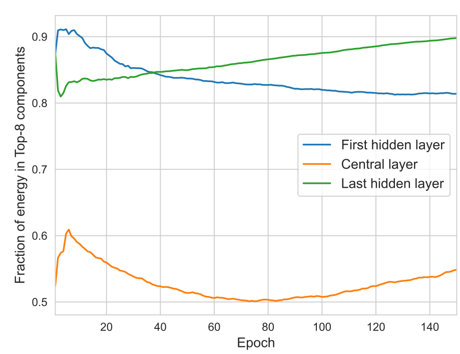

Figure 6: Fraction of energy in the top 8 components of the gradients of three layers in the network: the first and last hidden layer, and one central layer.

H.2 Static vs. dynamic power policy

As discussed in Sec. 4.2, we analyzed different power allocation schemes across iterations, when a fixed budget in terms of average power over the epochs is given. Fig. 3 shows the results for decreasing power allocations, while Fig. 7 here shows their increasing counterparts. We observe that LASER exhibits similar gains over Z-SGD for all the power control laws. Further, constant power remains the best policy for both LASER and Z-SGD. Whilst matching the constant power performance, the power-decreasing control performs better than the increasing counterpart for Z-SGD, especially in the low-power regime, where the accuracy gains are roughly $4 - 5\%$.

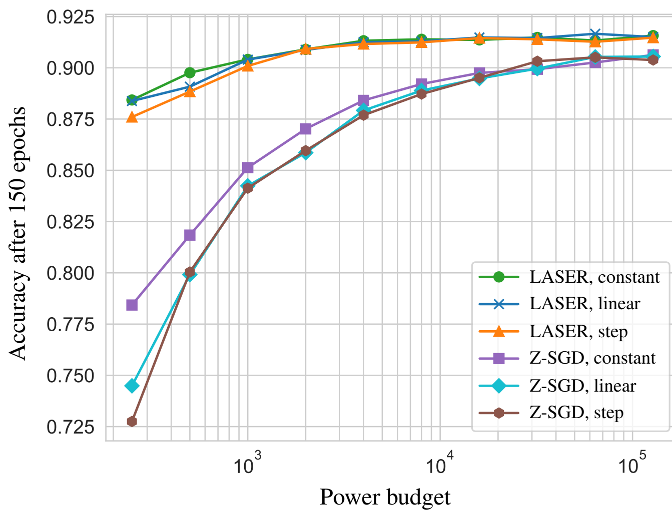

Figure 7: Final accuracy vs. power budget P with various power control schemes, for distributed training across 16 workers with ResNet18 on Cifar10. For each budget P, we consider three increasing power control laws, as studied in the literature [1], that satisfy the average power constraint: (i) constant power, $P_t = P$, (ii) piecewise constant, with the power levels $P_t \in \{P/3, 2P/3, P, 4P/3, 5P/3\}$, and (iii) linear law between the levels $P/3$ and $5P/3$. The performance of increasing power allocation schemes is equal or worse compared to their decreasing counterparts of Fig. 3.

H.3 Computational complexity and communication cost

Recall from Algorithm 1 that the two critical components of LASER are gradient compression and channel transmission. To gauge their efficacy we analyze them via two important metrics: (i) *computational complexity* of compression and (ii) *communication cost* of transmission. For (ii), recall from Eq. (1) that the power constraint indirectly serves as a communication cost and encourages compression. Table 3 quantitatively measures the total data sent by clients for each training iteration (doesn't change with the power P) for GPT language modeling on WIKITEXT-103. As illustrated, LASER incurs the lowest communication cost among all the baselines with $165\times$ cost reduction as compared to the Z-SGD, followed by SIGNUM which obtains $33\times$ reduction. Interestingly, LASER also achieves the best perplexity scores as highlighted in Fig. 1. For these experiments, we let rank $r = 4$ for LASER and the best compression factor 0.2 for the baselines (as detailed earlier). SIGNUM does not require any compression factor. For (i), since LASER relies on PowerSGD for the rank decomposition, it inherits the same low-complexity benefits: Tables 3-7 of [12] demonstrate that PowerSGD is efficient with significantly lower computational needs and has much smaller processing time/batch as compared to baselines without any accuracy drop. In fact, it is the core distributed algorithm behind the recent breakthrough DALL-E (§ E in [54]).

H.4 Slow and fast fading channels

The slow/non-fading model in Eq. (1) readily generalizes to the popular fast fading channel [13, 19]: $\boldsymbol{y} = \sum_i \gamma_i \boldsymbol{x}_i + \boldsymbol{Z}$, where γ_i are the channel fading coefficients. A standard technique here in the literature is to assume that channel-state-information (CSI) is known in the form of fading coefficients or their statistics, which essentially reduces the problem to a non-fading one. Likewise LASER can be extended to the fast fading channel as well. The challenging setting without CSI is an interesting topic of future research.