Organization. The appendix is organized as follows:

- App. A contains the requisite material for error feedback and SGD convergence analysis.
- App. B details the important technical lemmas needed for the theoretical convergence of LASER.
- App. C provides the proof for Theorem 1 whereasApp. D contains the proofs of all technical lemmas.
- App. E provides additional details about the noisy channel and Algorithm 1.
- App. F contains additional experimental details and results.

A 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 Stich & Karimireddy (2020). 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:

$$\widetilde{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \boldsymbol{e}_t. \tag{6}$$

Hence $\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t - \boldsymbol{e}_t - \gamma_t \boldsymbol{g}_t = \widetilde{\boldsymbol{\theta}}_t - \gamma_t \boldsymbol{g}_t$. 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 ((Stich & Karimireddy, 2020, 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.* (6)*,*

$$\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right) \mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - \frac{\gamma_t}{2}\mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star) + \gamma_t^2\sigma^2 + 3L\gamma_t\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2 . \tag{7}$$

The following lemma bounds the squared norm of the error, i.e. $\mathbb{E}\|\boldsymbol{e}_t\|^2$, appearing in Eq. (7). 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 (Stich & Karimireddy, 2020, Definition 10).

Lemma 2 ((Stich & Karimireddy, 2020, 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*

$$\mathbb{E}\left[3L\|\boldsymbol{e}_{t+1}\|^2\right] \leq \frac{\delta_r}{64L}\sum_{i=0}^{t}\left(1 - \frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \gamma_t\sigma^2 . \tag{8}$$

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 ((Stich & Karimireddy, 2020, 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 ((Stich & Karimireddy, 2020, 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*

- *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 $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$.

f is non-convex:

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 ((Stich & Karimireddy, 2020, 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.* (6)*,*

$$\mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_{t+1})] \leq \mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_t)] - \frac{\gamma_t}{4}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2}{2} + \frac{\gamma_t L^2}{2}\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2. \tag{9}$$

Lemma 5 ((Stich & Karimireddy, 2020, 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*

$$\mathbb{E}\left[3L\|\boldsymbol{e}_{t+1}\|^2\right] \leq \frac{\delta_r}{64L}\sum_{i=0}^{t}\left(1-\frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) + \gamma_t\sigma^2. \tag{10}$$

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 ((Stich & Karimireddy, 2020, 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 ((Stich & Karimireddy, 2020, 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 $\{\boldsymbol{\theta}_t\}_{t=0}^{T-1}$.

B 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

$$\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\underbrace{\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)}_{=\boldsymbol{P}\boldsymbol{Q}^\top}) \triangleq \sum_{i=1}^{r}\left(\boldsymbol{p}_i + \frac{\|\boldsymbol{p}_i\|}{\sqrt{\alpha_i}}\cdot \boldsymbol{Z}_m^{(i)}\right)\left(\boldsymbol{q}_i + \frac{\|\boldsymbol{q}_i\|}{\sqrt{\beta_i}}\cdot \boldsymbol{Z}_n^{(i)}\right)^\top, \tag{11}$$

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. B.1), how to control the influence of the channel $\mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}(\cdot)$ on the convergence of LASER (App. B.2), and utilize these results to establish technical lemmas along the lines of App. A for LASER (App. B.3).

B.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 $m \leq n$, 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 $S_P \triangleq \{(\alpha, \beta) : \alpha \geq 0, \beta \geq 0, \alpha + \beta = P\}$. Then for the minimizer $(\alpha^\star, \beta^\star) = \operatorname{argmin}_{(\alpha,\beta)\in S_P} f_P(\alpha, \beta)$, we have

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

Further the minimizer is given by

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

Lemma 8 (**Rank-r power allocation**). *For a power $P > 0$, $m, n, r \in \mathbb{N}$ with $m \leq n$, 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 $S_P \triangleq \{\{(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \boldsymbol{\alpha} \geq 0, \boldsymbol{\beta} \geq 0, \sum_i (\alpha_i + \beta_i) = P\}$. Then there exists a power allocation scheme $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) \in S_P$ such that

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

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

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

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.

B.2 Channel influence factor

In this section we establish the bounds for the channel influence defined in Eq. (4) 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*

$$\lambda_{\text{Z-SGD}} = \frac{1}{\mathrm{SNR}}, \tag{12}$$

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*

$$\lambda_{\text{LASER}} \leq \frac{4}{(m/r)\,\mathrm{SNR}}\left(1 + \frac{1}{(n/r)\,\mathrm{SNR}}\right), \tag{13}$$

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. D 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 Vogels et al. (2019) 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 $\lambda_{\text{LASER}} \leq \delta_r$.

B.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:

$$\widetilde{\boldsymbol{\theta}}_t \triangleq \boldsymbol{\theta}_t - \boldsymbol{e}_t\,. \tag{14}$$

Thus,

$$\widetilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_{t+1} - \boldsymbol{e}_{t+1} = \widetilde{\boldsymbol{\theta}}_t - \gamma_t \boldsymbol{g}_t + \mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t) - \mathcal{Z}_{(\boldsymbol{\alpha},\boldsymbol{\beta})}\left(\mathcal{C}_r(\boldsymbol{e}_t + \gamma_t \boldsymbol{g}_t)\right). \tag{15}$$

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.* (14)*,*

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}_\star\|^2 \leq \left(1 - \frac{\mu\gamma_t}{2}\right)\mathbb{E}\|\widetilde{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_\star\|^2 - \frac{\gamma_t}{2}\mathbb{E}(f(\boldsymbol{\theta}_t) - f_\star) + \gamma_t^2\sigma^2(1+\lambda_{\text{LASER}}) \\ + (3L\gamma_t(1+\lambda_{\text{LASER}}) + \lambda_{\text{LASER}})\mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2\,.\end{aligned} \tag{16}$$

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 $\boldsymbol{e}_t$ 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*

$$\begin{aligned}\left(3L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\mathbb{E}\|\boldsymbol{e}_{t+1}\|^2 \leq \frac{\delta_r}{32L}\sum_{i=0}^{t}\left(1 - \frac{\delta_r}{4}\right)^{t-i}\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_{t-i})\|^2\right) \\ + \gamma_t\sigma^2(1+\lambda_{\text{LASER}})\,.\end{aligned} \tag{17}$$

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

$$\begin{aligned}\left(3L(1+\lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right)\sum_{t=0}^{T} w_t\mathbb{E}\|\boldsymbol{e}_t\|^2 \leq \frac{1}{6L}\sum_{t=0}^{T} w_t\left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2\right) \\ + \sigma^2(1+\lambda_{\text{LASER}})\sum_{t=0}^{T} w_t\gamma_t\,.\end{aligned} \tag{18}$$

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.* (14)*,*

$$\begin{aligned}\mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_{t+1})] \leq \mathbb{E}[f(\widetilde{\boldsymbol{\theta}}_t)] - \frac{\gamma_t}{4}\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + \frac{\gamma_t^2 L\sigma^2(1+\lambda_{\text{LASER}})}{2} \\ + \mathbb{E}\|\boldsymbol{\theta}_t - \widetilde{\boldsymbol{\theta}}_t\|^2\left(\frac{L^2\gamma_t}{2} + L\lambda_{\text{LASER}}\right).\end{aligned} \tag{19}$$

C 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

$$\frac{w_t}{2} s_t \overset{(16)}{\leq} \frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + \gamma_t w_t \sigma^2 (1 + \lambda_{\text{LASER}}) + 3w_t (L(1 + \lambda_{\text{LASER}}) + \frac{\lambda_{\text{LASER}}}{\gamma_t}) \mathbb{E}\|\boldsymbol{e}_t\|^2 .$$

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

$$\sum_{t=0}^{T} \frac{w_t}{2} s_t \overset{(18)}{\leq} \sum_{t=0}^{T} \left(\frac{w_t}{\gamma_t}\left(1 - \frac{\mu\gamma_t}{2}\right) r_t - \frac{w_t}{\gamma_t} r_{t+1} + 2\gamma_t w_t \sigma^2 (1 + \lambda_{\text{LASER}}) \right) + \frac{1}{6L} \sum_{t=0}^{T} w_t \left(\mathbb{E}\|\nabla f(\boldsymbol{\theta}_t)\|^2 \right).$$

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 (Stich & Karimireddy, 2020, 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 $w_t = 1$.

(iii) f is non-convex 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

$$\frac{s_t}{4} \overset{(19)}{\leq} \frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8} + L\left(\frac{L}{2} + \frac{\lambda_{\text{LASER}}}{\gamma_t}\right) \mathbb{E}\|\boldsymbol{e}_t\|^2.$$

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

$$\frac{1}{4W_T} \sum_{t=0}^{T} w_t s_t \overset{(18)}{\leq} \frac{1}{W_T} \sum_{t=0}^{T} w_t \left(\frac{r_t}{4\gamma_t} - \frac{r_{t+1}}{4\gamma_t} + \frac{\gamma_t c}{8} \right) + \frac{L}{W_T} \left(\sum_{t=0}^{T} \frac{w_t s_t}{6L} + \frac{c w_t \gamma_t}{4L} \right),$$

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).

Z-SGD: 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

$$\begin{aligned}
\mathbb{E}\left[\mathcal{Z}_P(\boldsymbol{g}_t)|\boldsymbol{\theta}_t\right] &= \mathbb{E}[\boldsymbol{g}_t|\boldsymbol{\theta}_t] = \nabla f(\boldsymbol{\theta}_t), \\
\mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right] &= \mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \boldsymbol{g}_t + \boldsymbol{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right] \\
&= \mathbb{E}\left[\|\mathcal{Z}_P(\boldsymbol{g}_t) - \boldsymbol{g}_t\|^2|\boldsymbol{\theta}_t\right] + \mathbb{E}\left[\|\boldsymbol{g}_t - \nabla f(\boldsymbol{\theta}_t)\|^2|\boldsymbol{\theta}_t\right] \\
&\overset{4}{=} \mathbb{E}\left[\lambda_{\text{Z-SGD}}\|\boldsymbol{g}_t\|^2|\boldsymbol{\theta}_t\right] + \mathbb{E}\|\boldsymbol{\xi}_t\|^2 \\
&= \lambda_{\text{Z-SGD}}\|\nabla f(\boldsymbol{\theta}_t)\|^2 + (1 + \lambda_{\text{Z-SGD}})\mathbb{E}\|\boldsymbol{\xi}_t\|^2 \\
&\leq (M+1)(1 + \lambda_{\text{Z-SGD}})\|\nabla f(\boldsymbol{\theta}_t)\|^2 + (1 + \lambda_{\text{Z-SGD}})\sigma^2.
\end{aligned}$$

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}.

□

D PROOF OF TECHNICAL LEMMAS

D.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

$$\begin{aligned}
\alpha^\star &= \sqrt{1 + \frac{P}{n}}\left(\frac{\sqrt{1 + \frac{P}{m}} - \sqrt{1 + \frac{P}{n}}}{\frac{1}{m} - \frac{1}{n}}\right), \\
\beta^\star &= P - \alpha^\star.
\end{aligned} \tag{20}$$

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. (20) in terms of $\text{SNR} = P/mn$, we obtain

$$\frac{\alpha}{mn} = \frac{\sqrt{(1 + n\,\text{SNR})(1 + m\,\text{SNR})} - (1 + m\,\text{SNR})}{n - m}. \tag{21}$$

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. (21) as a function of γ. In particular, define $g(\gamma) \triangleq \sqrt{1+n\,\mathrm{SNR}}\sqrt{1+n\gamma\,\mathrm{SNR}}$. Observe that $g(1) = 1 + n\,\mathrm{SNR}$ and $g'(1) = \frac{n\,\mathrm{SNR}}{2}$. Rewriting Eq. (21) as a function of γ, we get

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

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:

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

□

D.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 $\sum_i \kappa_i = 1$. □

D.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

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

□

D.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. □

D.5 PROOF OF LEMMA 11

Proof. From Eq. (15), 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 $\mathbb{E}_{\boldsymbol{Z}}\|\text{Error}_{\boldsymbol{Z}}\|^2 \le \lambda_{\text{LASER}}\|\mathcal{C}_{\text{r}}(\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}})\|^2 \le \lambda_{\text{LASER}}\|\boldsymbol{e}_{\text{t}} + \gamma_{\text{t}}\boldsymbol{g}_{\text{t}}\|^2$ (see App. D.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 \\
&\le \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)}{\le} \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}}) \\
&\quad + (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 (22)
\end{aligned}$$

Now we closely follow the steps as in the proof of (Stich & Karimireddy, 2020, 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. (22), 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. □

D.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,

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

□

D.7 Proof of Lemma 13

Proof. From Eq. (15), 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. D.4). Using the smoothness of f,

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

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. □

E Additional details about noisy channel and LASER

E.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.

E.2 Detailed steps for Algorithm 1

Recall from Algorithm 1 that power allocation among clients is done via the function PowerAlloc($\{\mathcal{C}_r(\boldsymbol{M}_j), \boldsymbol{M}_j\}$). The theoretically optimal power allocation is discussed in App. B.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. F.6.

E.3 CONSTANT-ORDER SNR

As discussed in Sec. 3.2 and established in Lemmas 9 and 10 of App. B.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 Vogels et al. (2019) 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 $\lambda_{\text{Z-SGD}}$. 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).

F EXPERIMENTAL DETAILS

We provide technical details for the experiments demonstrated in Sec. 4.

F.1 WIKITEXT-103 EXPERIMENTAL SETUP

This section concerns the experimental details used to obtain Fig. 1 and Table 1 in the main text. Table 6 collects the settings we adopted to run our code. Table 7 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 6: Default experimental settings for the GPT-2 model used to learn the WIKITEXT-103 task.

Dataset	WIKITEXT-103
Architecture	GPT-2 (as implemented in Pagliardini (2023))
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 7: 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

F.2 CIFAR10 EXPERIMENTAL SETUP

This section concerns the experimental details used to obtain Fig. 2 and Table 3 in the main text. Table 8 collects the settings we adopted to run our code. Table 9 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 8: 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 9: 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

F.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 12 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 10 collects the settings we adopted to run our code. Table 11 describes the model architecture, with its parameters, their shape and their uncompressed size.

Table 10: 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 11: 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

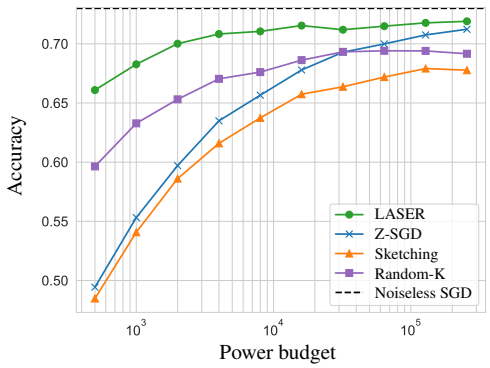

Figure 4: Test accuracy (*higher the better*) for a given power budget on CIFAR-100 for different algorithms. The advantage of LASER is evident across the entire power spectrum.

Table 12: 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	$32\times$

F.4 MNIST EXPERIMENTAL SETUP

This section concerns the experimental details used to obtain Table 4 in the main text. Table 13 collects the settings we adopted to run our code.

Table 13: 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

F.5 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.

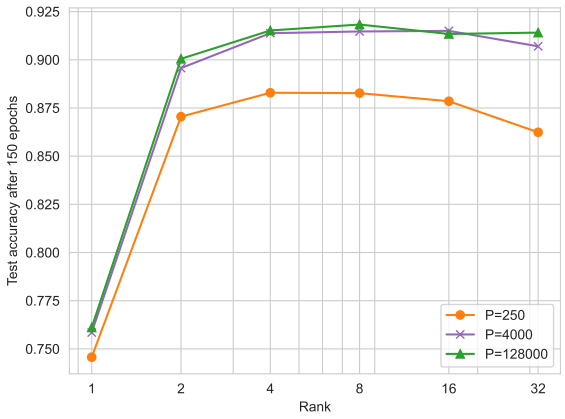

Figure 5: Final accuracy vs. compression rank tradeoff for CIFAR-10 classification, for low, medium and high power regimes. Rank-4/Rank-8 compression is optimal for all the three regimes. It reveals two interesting insights: (i) performance is uniformly worse in all the regimes with overly aggressive 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 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 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 Vogels et al. (2019), 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.

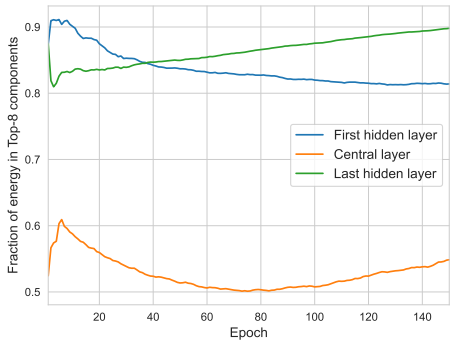

Figure 6: Fraction of energy in the top 8 components of the gradients of three layers in the network: the first and last hidden layer, and one central layer.

F.6 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. B.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.

F.7 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 $4-5\%$.

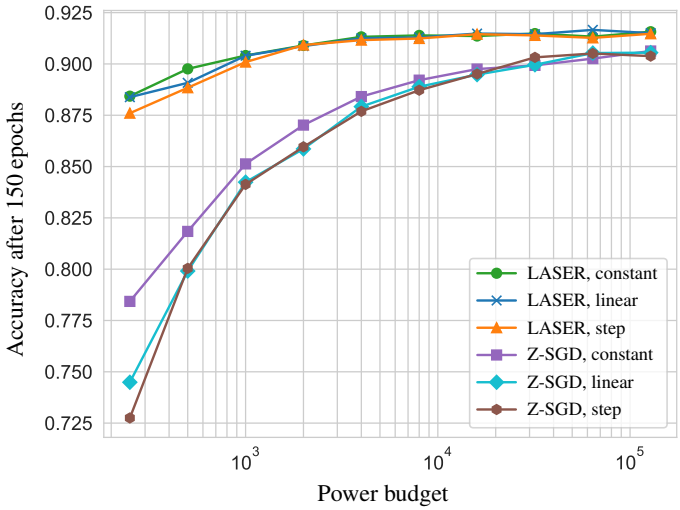

Figure 7: Final accuracy vs. power budget P with various power control schemes, for distributed training across 16 workers with RESNET18 on CIFAR10. For each budget P, we consider three increasing power control laws, as studied in the literature [1], that satisfy the average power constraint: (i) constant 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.

F.8 BASELINES IMPLEMENTATION

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

F.8.1 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.

F.8.2 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 $S_1 \ldots S_k$)
  $\hat{M} \leftarrow \mathbf{0} \in \mathbb{R}^{n\times m}$
  $\hat{M}(I) \leftarrow \frac{1}{k}\sum_{i=1}^{k} S_i$
  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.

F.8.3 SIGNUM

Algorithm 4 SIGNUM

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

We implemented SIGNUM following Bernstein et al. (2018). 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.