Basis Matters: Better Communication-Efficient Second Order Methods for Federated Learning

Xun Qian[1,2] Rustem Islamov[1,3] Mher Safaryan[1] Peter Richtárik[1]

[1]KAUST [2]JD Explore Academy [3]Institut Polytechnique de Paris

The Problem

$$\min_{x\in\mathbb{R}^d} f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where each function $f_i:\mathbb{R}^d\to\mathbb{R}$ represents the local loss associated with the data owned by device or client $i\in[n]\overset{\text{def}}{=}\{1,2,\dots,n\}$ only.

Algorithm

- $[\cdot]_\mu$: the projection on the set $\{\mathbf{A}\in\mathbb{R}^{d\times d} \mid \mathbf{A}=\mathbf{A}^\top, \mathbf{A}\succeq\mu\mathbf{I}\}$.

Algorithm 1: Basis Learn with Bidirectional Compression (BL1)

Parameters: Hessian learning rate $\alpha\geqslant 0$; model learning rate $\eta\geqslant 0$; gradient compression probability $p\in(0,1]$; compression operators $\{\mathcal{C}_1^k,\dots,\mathcal{C}_n^k\}$ and $\mathcal{Q}^k$; Basis $\{\mathbf{B}_i^{jl}\}$ in $\mathbb{R}^{d\times d}$ for each i

Initialization: $x^0=w^0=z^0\in\mathbb{R}^d$; $\mathbf{L}_i^0\in\mathbb{R}^{d\times d}$, $\mathbf{H}_i^0=\sum_{jl}(\mathbf{L}_i^0)_{jl}\mathbf{B}_i^{jl}$, and $\mathbf{H}^0\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\mathbf{H}_i^0$; $\xi^0=1$

for *each device* $i=1,\dots,n$ *in parallel* **do**

 if $\xi^k=1$

  $w^{k+1}=z^k$, compute $\nabla f_i(z^k)$ and send to the server

 if $\xi^k=0$

  $w^{k+1}=w^k$

 Compute $\nabla^2 f_i(z^k)$ and send $\mathbf{S}_i^k\overset{\text{def}}{=}\mathcal{C}_i^k(h^i(\nabla^2 f_i(z^k))-\mathbf{L}_i^k)$ to the server.

 Update local Hessian shifts $\mathbf{L}_i^{k+1}=\mathbf{L}_i^k+\alpha\mathbf{S}_i^k$,

 $\mathbf{H}_i^{k+1}=\mathbf{H}_i^k+\alpha\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$

end

on server

 if $\xi^k=1$

  $w^{k+1}=z^k,\ g^k=\nabla f(z^k)$

 if $\xi^k=0$

  $w^{k+1}=w^k,\ g^k=\left[\mathbf{H}^k\right]_\mu(z^k-w^k)+\nabla f(w^k)$

 $x^{k+1}=z^k-\left[\mathbf{H}^k\right]_\mu^{-1}g^k \quad \mathbf{H}^{k+1}=\mathbf{H}^k+\frac{\alpha}{n}\sum_{i=1}^n\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$

 Send $v^k\overset{\text{def}}{=}\mathcal{Q}^k(x^{k+1}-z^k)$ to all devices $i\in[n]$

 Update the model $z^{k+1}=z^k+\eta v^k$

 Send $\xi^{k+1}\sim\text{Bernoulli}(p)$ to all devices $i\in[n]$

for *each device* $i=1,\dots,n$ *in parallel* **do**

 Update the model $z^{k+1}=z^k+\eta v^k$

end

Compressor

- $\mathcal{C}:\mathbb{R}^{d\times d}\to\mathbb{R}^{d\times d}$ is called a *contraction compressor* if there exists a constant $0<\delta\le 1$ such that
$$\mathbb{E}\left[\|\mathbf{A}-\mathcal{C}(\mathbf{A})\|_{\mathrm{F}}^2\right]\le(1-\delta)\|\mathbf{A}\|_{\mathrm{F}}^2,\quad\forall\mathbf{A}\in\mathbb{R}^{d\times d}. \tag{2}$$
- $\mathcal{C}:\mathbb{R}^{d\times d}\to\mathbb{R}^{d\times d}$ is an *unbiased compressor* if there exists a constant $\omega\ge 0$ such that for any $\mathbf{A}\in\mathbb{R}^{d\times d}$
$$\mathbb{E}\left[\mathcal{C}(\mathbf{A})\right]=\mathbf{A}\text{ and }\mathbb{E}\left[\|\mathcal{C}(\mathbf{A})\|_{\mathrm{F}}^2\right]\le(\omega+1)\|\mathbf{A}\|_{\mathrm{F}}^2. \tag{3}$$

Assumptions

Assumption 1 (i)$\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is an unbiased compressor with parameter ω_{M} and $0<\eta\le 1/(\omega_{\mathrm{M}}+1)$. (ii) For all $j\in[d]$, $(z^k)_j$ in Algorithm 1 ($(z_i^k)_j$ in Algorithm 2) is a convex combination of $\{(x^t)_j\}_{t=0}^k$ for $k\ge 0$.

Assumption 2 (i) $\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is a contraction compressor with parameter δ_{M} and $\eta=1$. (ii) $\mathcal{Q}^k$ ($\mathcal{Q}_i^k$) is deterministic, i.e., $\mathbb{E}[\mathcal{Q}^k(x)]=\mathcal{Q}^k(x)$ for any $x\in\mathbb{R}^d$.

Assumption 3 (i) $\mathcal{C}_i^k$ is an unbiased compressor with parameter ω and $0<\alpha\le 1/(\omega+1)$.
(ii) For all $i\in[n]$ and $j,l\in[d]$, $(\mathbf{L}_i^k)_{jl}$ is a convex combination of $\{h^i(\nabla^2 f_i(z^t))_{jl}\}_{t=0}^k$ in Algorithm 1 ($\{h^i(\nabla^2 f_i(z_i^t))_{jl}\}_{t=0}^k$ in Algorithm 2) for $k\ge 0$.

Assumption 4 (i) $\mathcal{C}_i^k$ is a contraction compressor with parameter δ and $\alpha=1$. (ii) $\mathcal{C}_i^k$ is deterministic, i.e., $\mathbb{E}[\mathcal{C}_i^k(\mathbf{A})]=\mathcal{C}_i^k(\mathbf{A})$ for any $\mathbf{A}\in\mathbb{R}^{d\times d}$.

Assumption 5 We have $\|\nabla^2 f_i(x)-\nabla^2 f_i(y)\|\le H\|x-y\|$, $\|\nabla^2 f_i(x)-\nabla^2 f_i(y)\|_{\mathrm{F}}\le H_1\|x-y\|$, $\|h^i(\nabla^2 f_i(x))-h^i(\nabla^2 f_i(y))\|_{\mathrm{F}}\le M_1\|x-y\|$, $\max_{jl}\{|h^i(\nabla^2 f_i(x))_{jl}-h^i(\nabla^2 f_i(y))_{jl}|\}\le M_2\|x-y\|$, $\max_{jl}\{\|\mathbf{B}_i^{jl}\|_{\mathrm{F}}\}\le R$ for any $x,y\in\mathbb{R}^d$ and $i\in[n]$. For Algorithm 2, we assume each f_i is μ-strongly convex.

Some Notations

$$N_{\mathrm{B}}\overset{\text{def}}{=}\begin{cases}1 & \text{if the bases }\{\mathbf{B}_i^{jl}\}_{j,l\in[d]}\text{ are all orthogonal}\\ d^2 & \text{otherwise}\end{cases} \tag{4}$$

$$(A_{\mathrm{M}},B_{\mathrm{M}})\overset{\text{def}}{=}\begin{cases}(\eta,\eta) & \text{if Asm. 1(i) holds}\\ \left(\frac{\delta_{\mathrm{M}}}{4},\frac{6}{\delta_{\mathrm{M}}}-\frac{7}{2}\right) & \text{if Asm. 2(i) holds}\end{cases} \tag{5}$$

$$(A,B)\overset{\text{def}}{=}\begin{cases}(\alpha,\alpha) & \text{if Asm. 3(i) holds}\\ \left(\frac{\delta}{4},\frac{6}{\delta}-\frac{7}{2}\right) & \text{if Asm. 4(i) holds}\end{cases} \tag{6}$$

For any $k\ge 0$, denote $\mathcal{H}^k\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\|\mathbf{L}_i^k-\mathbf{L}_i^*\|_{\mathrm{F}}^2$, $\Phi_1^k\overset{\text{def}}{=}\|z^k-x^*\|^2+\frac{A_{\mathrm{M}}(1-p)}{2p}\|w^k-x^*\|^2$, where $\mathbf{L}_i^*\overset{\text{def}}{=}h^i(\nabla^2 f_i(x^*))$, $\Phi_2^k\overset{\text{def}}{=}\mathcal{H}^k+\frac{4BM_1^2}{A_{\mathrm{M}}}\|x^k-x^*\|^2$.

Convergence Result

Linear convergence of BL1: Let Assumption 5 hold. Let Assumption 1 (i) or Assumption 2 (i) hold. Assume $\|z^k-x^*\|^2\le\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}$ and $\mathcal{H}^k\le\frac{A_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for $k\ge 0$. Thenxfor $k\ge 0$ we have

$$\mathbb{E}[\Phi_1^k]\le\left(1-\frac{\min\{A_{\mathrm{M}},p\}}{2}\right)^k\Phi_1^0.$$

Superlinear convergence of BL1: Let $\eta=1$, $\xi^k\equiv 1$ and $\mathcal{Q}^k(x)\equiv x$ for any $x\in\mathbb{R}^d$ and $k\ge 0$. Let Assumption 5 hold. Let Assumption 3 (i) or Assumption 4 (i) hold. Assume $\|z^k-x^*\|^2\le\frac{A_{\mathrm{M}}\mu^2}{4H^2B_{\mathrm{M}}}$ and $\mathcal{H}^k\le\frac{A_{\mathrm{M}}\mu^2}{16N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for $k\ge 0$. Then we have

$$\mathbb{E}[\Phi_2^k]\le\theta_1^k\Phi_2^0,$$

$$\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right]\le\theta_1^k\left(\frac{A_{\mathrm{M}}H^2}{8BM_1^2\mu^2}+\frac{2N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_2^0,$$

for $k\ge 0$, where $\theta_1\overset{\text{def}}{=}\left(1-\frac{\min\{4A,A_{\mathrm{M}}\}}{4}\right)$.

Algorithm

- $[\mathbf{A}]_s=(\mathbf{A}+\mathbf{A}^\top)/2$ for any $\mathbf{A}\in\mathbb{R}^{d\times d}$.

Algorithm 2: Basis Learn with Bidirectional Compression and Partial Participation (BL2)

Parameters: $\alpha>0$; $\eta>0$; matrix compression operators $\{\mathcal{C}_1^k,\dots,\mathcal{C}_n^k\}$; $p\in(0,1]$; $0<\tau\le n$

Initialization: $w_i^0=z_i^0=x^0\in\mathbb{R}^d$; $\mathbf{L}_i^0\in\mathbb{R}^{d\times d}$; $\mathbf{H}_i^0=\sum_{jl}(\mathbf{L}_i^0)_{jl}\mathbf{B}_i^{jl}$; $l_i^0=\|[\mathbf{H}_i^0]_s-\nabla^2 f_i(w_i^0)\|_{\mathrm{F}}$; $g_i^0=([\mathbf{H}_i^0]_s+l_i^0\mathbf{I})w_i^0-\nabla f_i(w_i^0)$; Moreover: $\mathbf{H}^0=\frac{1}{n}\sum_{i=1}^n\mathbf{H}_i^0$; $l^0=\frac{1}{n}\sum_{i=1}^n l_i^0$; $g^0=\frac{1}{n}\sum_{i=1}^n g_i^0$

on server

 $x^{k+1}=\left([\mathbf{H}^k]_s+l^k\mathbf{I}\right)^{-1}g^k$, choose a subset $S^k\subseteq[n]$ such that $\mathbb{P}[i\in S^k]=\tau/n$ for all $i\in[n]$

 $v_i^k=\mathcal{Q}_i^k(x^{k+1}-z_i^k)$, $z_i^{k+1}=z_i^k+\eta v_i^k$ for $i\in S^k$

 $z_i^{k+1}=z_i^k,\quad w_i^{k+1}=w_i^k$ for $i\notin S^k$

 Send v_i^k to the selected devices $i\in S^k$

for *each device* $i=1,\dots,n$ *in parallel* **do**

 for participating devices $i\in S^k$ **do**

 $z_i^{k+1}=z_i^k+\eta v_i^k$, $\mathbf{S}_i^k\overset{\text{def}}{=}\mathcal{C}_i^k(h^i(\nabla^2 f_i(z_i^{k+1}))-\mathbf{L}_i^k)$

 $\mathbf{L}_i^{k+1}=\mathbf{L}_i^k+\alpha\mathbf{S}_i^k$, $\mathbf{H}_i^{k+1}=\mathbf{H}_i^k+\alpha\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$

 $l_i^{k+1}=\|[\mathbf{H}_i^{k+1}]_s-\nabla^2 f_i(z_i^{k+1})\|_{\mathrm{F}}$

 Sample $\xi_i^{k+1}\sim\text{Bernoulli}(p)$

 if $\xi_i^k=1$

  $w_i^{k+1}=z_i^{k+1}$, $g_i^{k+1}=([\mathbf{H}_i^{k+1}]_s+l_i^{k+1}\mathbf{I})w_i^{k+1}-\nabla f_i(w_i^{k+1})$, send $g_i^{k+1}-g_i^k$ to server

 if $\xi_i^k=0$

  $w_i^{k+1}=w_i^k$, $g_i^{k+1}=([\mathbf{H}_i^{k+1}]_s+l_i^{k+1}\mathbf{I})w_i^{k+1}-\nabla f_i(w_i^{k+1})$

 Send $\mathbf{S}_i^k$, $l_i^{k+1}-l_i^k$, and ξ_i^k to server

 for non-participating devices $i\notin S^k$ **do**

 $z_i^{k+1}=z_i^k$, $w_i^{k+1}=w_i^k$, $\mathbf{L}_i^{k+1}=\mathbf{L}_i^k$, $\mathbf{H}_i^{k+1}=\mathbf{H}_i^k$, $l_i^{k+1}=l_i^k$, $g_i^{k+1}=g_i^k$

end

on server

 if $\xi_i^k=1$

  $w_i^{k+1}=z_i^{k+1}$, receive $g_i^{k+1}-g_i^k$

 if $\xi_i^k=0$

  $w_i^{k+1}=w_i^k$, $g_i^{k+1}-g_i^k=\alpha\left[\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}\right]_s w_i^{k+1}+(l_i^{k+1}-l_i^k)w_i^{k+1}$

 $g^{k+1}=g^k+\frac{1}{n}\sum_{i\in S^k}\left(g_i^{k+1}-g_i^k\right)$

 $\mathbf{H}^{k+1}=\mathbf{H}^k+\frac{\alpha}{n}\sum_{i\in S^k}\sum_{jl}(\mathbf{S}_i^k)_{jl}\mathbf{B}_i^{jl}$

 $l^{k+1}=l^k+\frac{1}{n}\sum_{i\in S^k}\left(l_i^{k+1}-l_i^k\right)$

Convergence Result

Let $\Phi_3^k\overset{\text{def}}{=}\mathcal{W}^k+\frac{2p}{A_{\mathrm{M}}}\left(1-\frac{\tau A_{\mathrm{M}}}{n}\right)\mathcal{Z}^k$, where $\mathcal{Z}^k\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\|z_i^k-x^*\|^2$, for $k\ge 0$.

Linear convergence of BL2: Let Assumption 5 hold. Let Assumption 1 (i) or Assumption 2 (i) hold. Assume $\|z_i^k-x^*\|^2\le\frac{A_{\mathrm{M}}\mu^2}{(6H^2+24H_1^2)B_{\mathrm{M}}}$ and $\mathcal{H}^k\le\frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for all $i\in[n]$ and $k\ge 0$. Then for $k\ge 0$

$$\mathbb{E}[\Phi_3^k]\le\left(1-\frac{\tau\min\{p,A_{\mathrm{M}}\}}{2n}\right)^k\Phi_3^0.$$

Superlinear convergence of BL2: Define $\Phi_4^k\overset{\text{def}}{=}\mathcal{H}^k+\frac{4BM_1^2}{A_{\mathrm{M}}}\|x^k-x^*\|^2$ for $k\ge 0$. Let $\eta=1$, $\xi^k\equiv 1$, $S^k\equiv[n]$, and $\mathcal{Q}_i^k(x)\equiv x$ for any $x\in\mathbb{R}^d$ and $k\ge 0$. Let Assumption 5 hold. Let Assumption 3 (i) or Assumption 4 (i) hold. Assume $\|z_i^k-x^*\|^2\le\frac{A_{\mathrm{M}}\mu^2}{(6H^2+24H_1^2)B_{\mathrm{M}}}$ and $\mathcal{H}^k\le\frac{A_{\mathrm{M}}\mu^2}{96N_{\mathrm{B}}R^2B_{\mathrm{M}}}$ for all $i\in[n]$ and $k\ge 0$. Then we have

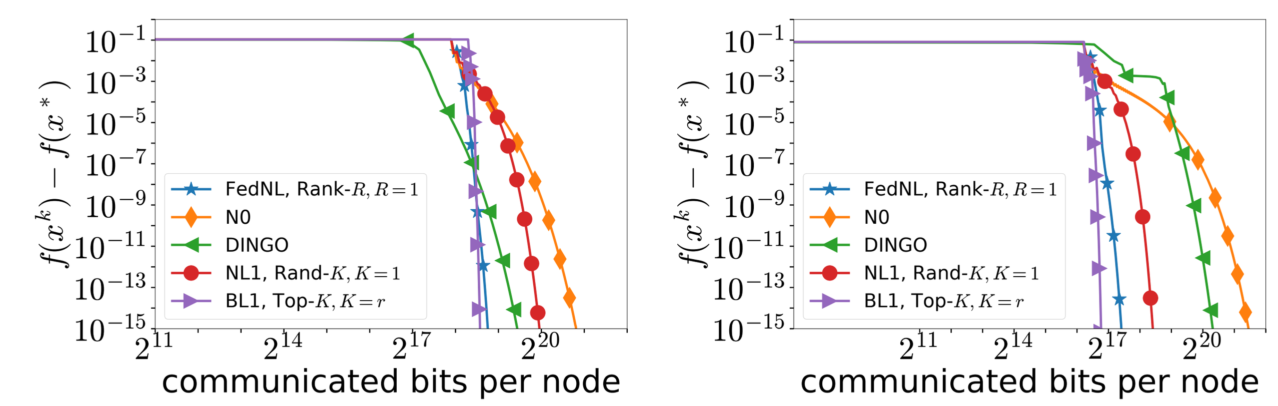
$$\mathbb{E}[\Phi_4^k]\le\theta_2^k\Phi_4^0,$$

$$\mathbb{E}\left[\frac{\|x^{k+1}-x^*\|^2}{\|x^k-x^*\|^2}\right]\le\theta_2^k\left(\frac{A_{\mathrm{M}}(3H^2+12H_1^2)}{16BM_1^2\mu^2}+\frac{12N_{\mathrm{B}}R^2}{\mu^2}\right)\Phi_4^0,$$

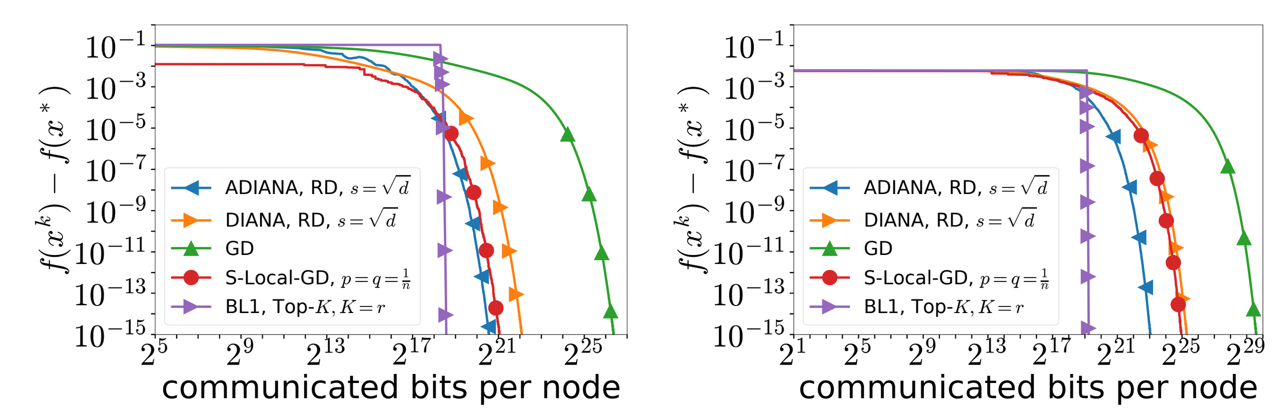
for $k\ge 0$, where $\theta_2\overset{\text{def}}{=}\left(1-\frac{\min\{2A,A_{\mathrm{M}}\}}{2}\right)$.

Numerical Results

1. BL1 vs N0 vs FedNL vs NL1 vs DINGO



2. BL1 vs DIANA vs ADIANA vs GD vs S-Local-GD



3. ECLK vs ADIANA