Optimization for Machine Learning

Lecture 12: Compression (with Error-Feedback)

Sebastian Stich

CISPA – `https://cms.cispa.saarland/optml24/`

July 9, 2024

Projects

Project: Final steps

- Poster printing: please send your poster in pdf format to Yuan Gao (yuan.gao@cispa.de) before Monday, July 15, **8am**.
- (You can also print the poster yourself. We can reimburse the costs up to 20 EUR in exchange of a proper receipt.)
- Upload the final report by June 26 to CMS (you can make adjustments after the poster presentation, and take suggestions/comments into account).

Lecture: July 16

- 16:15h, Research Talk by Kumar Kshitij Patel, (PhD Student at TTIC).
- 17:15-18:00h, Poster Session.

Exam Factsheet

- 2.5 hours
- closed book
  - you can bring **one double-sided A4 page** cheat sheet
- materials
  - all topics covered in the lecture
- practice exams
  - link to old exams posted on the course website
  - note that for these exams the syllabus might have been (slightly) different

Exam Registration (on CMS/LSF)

- mandatory, latest 1 week before the exam!
- please register early, the deadline is strict even if there are technical problems (on either side)
- the registration link should work for all that have finalized their project
- if you cannot register (but think you should be able to) please reach out ASAP!

Evaluation (UdS)

Please fill out the evaluation forms provided by UdS:

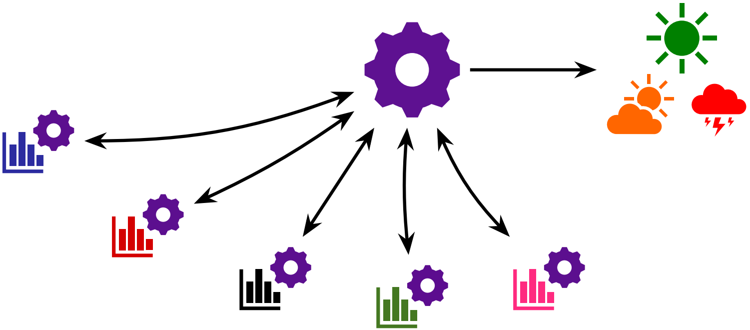
Lecture: Link to the Evaluation form for the Lecture

Exercises: Link to eh Evaluation form for the Exercises

(you can click on these links, or you find the same link also on the course material page)

Lecture 12

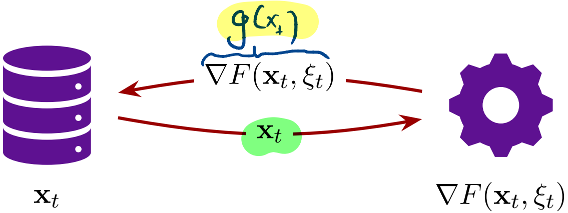
Distributed Training



- limited bandwidth connections
- high latency

Communication Bottleneck

SGD: $x_{t+1} = x_t - \gamma g(x_t)$



- We need to communicate $\mathbb{R}^d$ vectors (model parameters, or gradients) in every communication round.

Q: Can we compress these messages?

Lecture Outline

Training Objective

$$\min_{\mathbf{x}\in\mathbb{R}^d}\left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{f_i(\mathbf{x})}_{\text{data } \mathcal{D}_i \text{ on client } i}\right] \qquad f_i(\mathbf{x}) = \begin{cases} \mathbb{E}_{\xi\sim\mathcal{D}_i} F(\mathbf{x},\xi) \\ \frac{1}{m}\sum_{j=1}^{m} f_{ij}(\mathbf{x}) \end{cases}$$

- For simplicity, we will again first discuss the homogeneous setting $(f_i = f_j, \forall i, j)$.

Simplified Scenario:

- Consider $n = 1$ worker device, that communicates with a server.

Baseline: Stochastic Gradient Descent

Stochastic Gradient Descent (SGD): γ stepsize

$$\underbrace{\mathbf{g}_t = \mathbf{g}(\mathbf{x})}_{\text{uniform data sample}} \qquad \mathbf{x}_{t+1} := \underbrace{\mathbf{x}_t - \gamma \mathbf{g}_t}_{\text{model update}}$$

Assumptions:

- $f \colon \mathbb{R}^d \to \mathbb{R}$ convex and L-smooth
- $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x})$, $\forall x \in \mathbb{R}^d$
- $\mathbb{E}\left\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq \sigma^2$, $\forall x \in \mathbb{R}^d$

Convergence: the iteration complexity to reach $\mathbb{E} f(\mathbf{x}_{\text{out}}) - f^\star \leq \epsilon$ is

$$\mathcal{O}\left(\frac{\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right) \cdot R_0$$

with $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.

Lecture Outline

Motivation

- Instead of sending the full gradient vector $\mathbf{g}_t \in \mathbb{R}^d$ from the worker to the server, can we compress the gradient?

Schematic:

$$\underbrace{\mathbf{g}_t}_{\text{gradient}} \rightarrow \underbrace{\mathcal{Q}(\mathbf{g}_t)}_{\text{compression}} \underbrace{\rightarrow}_{\text{send to server}} \underbrace{\mathcal{Q}^{-1}(\mathcal{Q}(\mathbf{g}_t))}_{\text{de-compression}} \rightarrow \underbrace{\tilde{\mathbf{g}}_t}_{\text{approximate gradient}}$$

Compressor:

- $\mathcal{Q}$: $\mathbb{R}^d \rightarrow \mathcal{X}$ (possibly lossy compression)
- $\mathcal{Q}^{-1}$: $\mathcal{X} \rightarrow \mathbb{R}^d$

Convention:

- We will often use the shorthand $\mathcal{Q}(\mathbf{g})$ to denote $\mathcal{Q}^{-1}(\mathcal{Q}(\mathbf{g})) \in \mathbb{R}^d$.

Properties

Motivation:

- Suppose we want to study compressed SGD:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathcal{Q}(\mathbf{g}_t)$$

Some compression of g_t

- It would be very convenient if $\mathbb{E}[\mathcal{Q}(\mathbf{g})] = \nabla f(\mathbf{x})$.

Definition 12.1 (Unbiased ω-quantization)

A compressor $\mathcal{Q} \colon \mathbb{R}^d \to \mathbb{R}^d$ is an unbiased $\omega \geq 0$ quantizer, if

$$\mathbb{E}_{\mathcal{Q}} \mathcal{Q}(\mathbf{x}) = \mathbf{x}, \qquad \forall \mathbf{x} \in \mathbb{R}^d$$

unbiased

and

$$\mathbb{E}_{\mathcal{Q}} \left\| \mathcal{Q}(\mathbf{x}) - \mathbf{x} \right\|^2 \leq \omega \left\| \mathbf{x} \right\|^2, \qquad \forall \mathbf{x} \in \mathbb{R}^d.$$

Examples

$$x \in \mathbb{R}^d = \begin{pmatrix} x \\ \circ \\ x \\ \vdots \\ x \\ \circ \end{pmatrix} \Big\} k \text{ coordinates} \Rightarrow Q(x) : \begin{pmatrix} 0 \\ 0 \\ x \\ 0 \\ x \\ 0 \end{pmatrix} \cdot \underbrace{\frac{d}{k}}_{\text{to make it unbiased!}}$$

- random sparsification
  $\mathcal{Q}(\mathbf{x}) = \frac{d}{k} \cdot M \odot \mathbf{x}$, where $M \in \{0,1\}^d$ is a mask that selects k random coordinates

$$E[Q(x)]_i = \underbrace{\sum}_{\text{coordinate } i \text{ is picked or not}} \frac{k}{d} \cdot (x_i) \cdot \frac{d}{k} + \left(1 - \frac{k}{d}\right) \cdot 0 = (x_i)$$

Exercise: compute w!

- quantization
  $\mathcal{Q}(\mathbf{x}) = \operatorname{sign}(\mathbf{x}) \cdot \|\mathbf{x}\| \cdot \frac{1}{s} \cdot \operatorname{round}\left(s \frac{|\mathbf{x}|}{\|\mathbf{x}\|}\right)$,

$$\text{where } \operatorname{round}(x) = \begin{cases} \lceil x \rceil, \; (=2) & \text{with probability } x - \lfloor x \rfloor \quad 30\% \\ \lfloor x \rfloor \; (=1) & \text{with probability } \lceil x \rceil - x \quad 70\% \end{cases}$$

$x = 1.3$

Expectation: $0.7 \cdot 1 + 0.3 \cdot 2$
$= 1.3$

Quantized SGD [AGL+17]

Quantized SGD

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, ω-quantizer $\mathcal{Q}$, $\gamma > 0$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathcal{Q}(\mathbf{g}(\mathbf{x}))\,.$$

Theorem 12.2

Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be convex, L-smooth and let $R_0 = \left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2$ and $\gamma \leq \frac{1}{2L(1+\omega)}$. Then there exists a stepsize γ such that $\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E} f(\mathbf{x}_t) - f^\star\right) \leq \epsilon$ for

$$T = \mathcal{O}\left(\frac{\sigma^2}{\epsilon^2} + \frac{L}{\epsilon}\right) \cdot R_0 \cdot (1+\omega)$$

iterations of quantized SGD with an ω-quantizer.

↑ more iterations!

Proof

We expand, and use the property of the ω-quantizer:

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^\star\right\|^2 &= \mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2 - 2\gamma\mathbb{E}\mathcal{Q}(\mathbf{g}(\mathbf{x}_t))^\top(\mathbf{x}_t-\mathbf{x}^\star) + \gamma^2\mathbb{E}\left\|\mathcal{Q}(\mathbf{g}(\mathbf{x}_t))\right\|^2 \\
&\qquad\text{• unbiased quant!} \\
&\le \mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2 - 2\gamma\mathbb{E}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star) + \gamma^2\underline{(1+\omega)}\left(\left\|\nabla f(\mathbf{x}_t)\right\|^2+\sigma^2\right) \\
&\le \mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2 - 2\gamma(\mathbb{E}f(\mathbf{x}_t)-f^\star) + \gamma^2(1+\omega)\left(2L(\mathbb{E}f(\mathbf{x}_t)-f^\star)+\sigma^2\right)
\end{aligned}
$$

with convexity and smoothness. Now, by $\gamma \le \frac{1}{2(1+\omega)L}$,

$$
\gamma(\mathbb{E}f(\mathbf{x}_t)-f^\star) \le \mathbb{E}\left\|\mathbf{x}_t-\mathbf{x}^\star\right\|^2 - \mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}^\star\right\|^2 + \gamma^2(1+\omega)\sigma^2 \,.
$$

With the usual procedure, summing over $t=0,\ldots,T-1$, dividing by T and γ:

$$
\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E}f(\mathbf{x}_t)-f^\star) \le \frac{R_0}{\gamma} + \underline{\gamma(1+\omega)\sigma^2}
$$

and the theorem follows by minimizing in γ.

Discussion

- While quantization decreases the per-iteration communication cost, the worst-case complexity bounds to not show a total speedup, when taking the full cost of the optimization (iterations $\times$ cost per iteration) into account.
- In practice, a speedup can often be still observed.
- In practice, the 'top-k' compressor often significantly outperforms 'random-k' (which is supported by theory).

Q: Can we compressed SGD converge with a provable speedup, supporting also biased compressors such as 'top-k'?

Biased Compressors

Definition 12.3 ((biased) δ-compressor)

A compressor $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ is an $\delta > 0$ compressor, if

$$\mathbb{E}_{\mathcal{C}} \left\| \mathcal{C}(\mathbf{x}) - \mathbf{x} \right\|^2 \leq (1-\delta) \left\| \mathbf{x} \right\|^2 , \qquad \forall \mathbf{x} \in \mathbb{R}^d .$$

- Note that we do not impose a condition for unbiasedness.
- If $\mathcal{Q}(\mathbf{x})$ is a ω quantizer, then $\frac{1}{1+\omega}\mathcal{Q}(\mathbf{x})$ is a $\delta = \frac{1}{1+\omega}$ compressor (exercise).

Examples

- random sparsification
  $\mathcal{C}(\mathbf{x}) = M \odot \mathbf{x}$, where $M \in \{0,1\}^d$ is a mask that selects k random coordinates.

- top-k sparsification
  $\mathcal{C}(\mathbf{x}) = \mathrm{top}_k(\mathbf{x})$

$k=1$ pick index i, s.t. $|x_i| \geq |x_j| \ \forall j \neq i$

$\delta = \frac{1}{d}$, d dimension

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \rightsquigarrow \mathcal{C}(x) = \begin{pmatrix} 0 \\ 0 \\ x_i \\ 0 \\ 0 \end{pmatrix}$$

$$\|\mathrm{top}_1(x) - x\|^2 = \sum_{i=1}^{d} (\mathrm{top}_1(x) - x)_i^2 = \sum_{i \neq \text{top 1 coordinate}} (x)_i^2$$

$$\leq \left(1 - \frac{1}{d}\right) \cdot \sum_{i=1}^{d} (x_i)^2 = \left(1-\frac{1}{d}\right) \cdot \|x\|^2$$

- rank-k approximation
- arbitrary black box compressors: Zip, JPEG, etc.

Error Feedback SGD/Error Compensated SGD

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, stepsize $\gamma > 0$, correction buffer $\mathbf{e}_0 = \mathbf{0} \in \mathbb{R}^d$. At iteration t:

$$
\begin{aligned}
\mathbf{g}_t &= \mathbf{g}(\mathbf{x}_t) && \text{(stochastic gradient)} \\
\mathbf{v}_t &= \mathcal{C}(\mathbf{e}_t + \gamma \mathbf{g}_t) && \text{(compressed \& error compensated update)} \\
\mathbf{x}_{t+1} &= \mathbf{x}_t - \mathbf{v}_t && \\
\mathbf{e}_{t+1} &= \mathbf{e}_t + \underbrace{\gamma \mathbf{g}_t - \mathbf{v}_t}_{\text{Compression error}} && \text{(tracking the compession error)}
\end{aligned}
$$

Convergence

Theorem 12.4 ([SCJ18, SK20])

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, L-smooth and let $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$ and $\gamma \leq \frac{\delta}{10L}$. Then there exists a stepsize γ such that $\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E} f(\mathbf{x}_t) - f^\star) \leq \epsilon$ for

$$T = \mathcal{O}\left(\frac{\sigma^2}{\epsilon^2} + \underbrace{\frac{\sqrt{(1-\delta)L\sigma^2}}{\epsilon^{3/2}\delta}}_{\text{"ignore"}} + \frac{L}{\delta\epsilon}\right) \cdot R_0 \,,$$

iterations of error-compensated SGD with an δ-compressor.

- The compressor quality δ only impacts the optimization term, but not the stochastic term.
- For instance, for a compressor with $\delta = \frac{1}{1+\omega}$, the speedup can reach a factor of $(1+\omega)$ in comparison to quantization without error feedback (with the same per-iteration communication costs).

Convergence

- The proof of Theorem 12.4 follows a similar template as the proof for asynchronous SGD/Hogwild. However, as the technical details are somewhat more involved, we leave the full prove as an exercise and prove here a variant under stronger assumptions:

Theorem 12.5

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex, L-smooth and let $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$ and $\gamma \leq \frac{1}{4L}$. Additionally, assume the stochastic gradients are bounded, $\mathbb{E}\|\mathbf{g}_t\|^2 \leq B^2$, $\forall t$. Then there exists a stepsize γ such that $\frac{1}{T}\sum_{t=0}^{T-1}(\mathbb{E} f(\mathbf{x}_t) - f^\star) \leq \epsilon$ for

$$T = \mathcal{O}\left(\frac{B^2}{\epsilon^2} + \frac{\sqrt{(1-\delta)LB^2}}{\epsilon^{3/2}\delta} + \frac{L}{\epsilon}\right) \cdot R_0\,,$$

δ is hidden in "the assuption"

iterations of error-compensated SGD with an δ-compressor.

- Note: the strong condition $\mathbb{E}\|\mathbf{g}_t\|^2 \leq B^2$ allow us to relax the condition on the stepsize ($\gamma \leq \frac{\delta}{10L}$ in Theorem 4, vs. $\gamma \leq \frac{1}{4L}$ in Theorem 5).

Proof I: Virtual Sequence

For the analysis, it will be convenient to define a sequence of virtual iterates $\tilde{\mathbf{x}}_t$. We define

$$\tilde{\mathbf{x}}_t = \mathbf{x}_t - \mathbf{e}_t$$

with $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$ (note that $\mathbf{e}_0 = \mathbf{0}$). We observe that

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{x}_{t+1} - \mathbf{e}_{t+1} = (\mathbf{x}_t - \mathbf{v}_t) - (\mathbf{e}_t + \gamma \mathbf{g}_t - \mathbf{v}_t) = \tilde{\mathbf{x}}_t - \gamma \mathbf{g}_t .$$

Proof II: Technical Lemmas

Lemma 12.6 (Decrease)

For $\gamma \leq \gamma_{\text{crit}} = \frac{1}{4L}$ it holds

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2 \leq \mathbb{E}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 - \frac{\gamma}{2}(\mathbb{E}f(\mathbf{x}_t) - f^\star) + \gamma^2\sigma^2 + 2L\gamma\mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2$$

$$\left(\leq \mathbb{E}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 - \frac{\gamma}{2}(\mathbb{E}f(\mathbf{x}_t) - f^\star) + \gamma^2 B^2 + 2L\gamma\mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2 \right).$$

- *We use $\sigma^2 \leq B^2$. The first equation can also be used in the proof of Theorem 4 (see exercises).*

Lemma 12.7 (Difference)

With the notation for $R_t = \left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2$, it holds

$$\mathbb{E}R_t \leq \frac{4(1-\delta)\gamma^2 B^2}{\delta^2}$$

Proof III: Combine the Lemmas

We now plug Lemma 12.6 into the statement of Lemma 12.7:

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1}-\mathbf{x}^\star\right\|^2 \leq \mathbb{E}\left\|\tilde{\mathbf{x}}_t-\mathbf{x}^\star\right\|^2 - \frac{\gamma}{2}(\mathbb{E}f(\mathbf{x}_t)-f^\star) + \gamma^2 B^2 + \frac{8(1-\delta)L\gamma^3 B^2}{\delta^2}.$$

Now we re-arrange, sum over $t = 0, \ldots, T-1$ and divide by (γT):

$$\begin{aligned}\frac{1}{2T}\sum_{t=0}^{T-1}(\mathbb{E}f(\mathbf{x}_t)-f^\star) &\leq \frac{1}{\gamma T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left\|\tilde{\mathbf{x}}_t-\mathbf{x}^\star\right\|^2 - \mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1}-\mathbf{x}^\star\right\|^2\right) + \gamma B^2 + \frac{8(1-\delta)L\gamma^2 B^2}{\delta^2} \\ &= \mathcal{O}\left(\frac{R_0}{\gamma T} + \gamma B^2 + \gamma^2\frac{(1-\delta)LB^2}{\delta^2}\right).\end{aligned}$$

Now the proof follows by chosing the optimal stepsize (see Exercise Sheet 6).

Proof of Lemma 12.6

We prove here a stronger statement. We expand and take expectation:

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1}-\mathbf{x}^\star\right\|^2 = \mathbb{E}\left\|\tilde{\mathbf{x}}_{t}-\mathbf{x}^\star\right\|^2 - 2\gamma\mathbb{E}\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) + \gamma^2\mathbb{E}\left\|\mathbf{g}_t\right\|^2 + 2\gamma\mathbb{E}\mathbf{g}_t^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_t)$$
$$\leq \mathbb{E}\left\|\tilde{\mathbf{x}}_{t}-\mathbf{x}^\star\right\|^2 - 2\gamma\mathbb{E}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star) + \gamma^2(\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2+\sigma^2) + 2\gamma\mathbb{E}\nabla f(\mathbf{x}_t)^\top\underbrace{(\mathbf{x}_t-\tilde{\mathbf{x}}_t)}_{\text{difference}}$$

Now we use:

- $-\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star) \leq -(f(\mathbf{x}_t)-f^\star)$, by convexity
- $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\tilde{\mathbf{x}}_t) \leq \frac{1}{2L}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + 2L\left\|\mathbf{x}_t-\tilde{\mathbf{x}}_t\right\|^2$ $\qquad \mathbf{a}^\top\mathbf{b} \leq \frac{1}{2\lambda}\left\|\mathbf{a}\right\|^2 + \frac{\lambda}{2}\left\|\mathbf{b}\right\|^2$
- $\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq 2L(f(\mathbf{x}_t)-f^\star)$, by smoothness

Putting all these together:

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_{t+1}-\mathbf{x}^\star\right\|^2 \leq \mathbb{E}\left\|\tilde{\mathbf{x}}_{t}-\mathbf{x}^\star\right\|^2 - \gamma\left(2-2L\gamma-1\right)\left(\mathbb{E}f(\mathbf{x}_t)-f^\star\right) + \gamma^2\sigma^2 + 2L\gamma\mathbb{E}\left\|\mathbf{x}_t-\tilde{\mathbf{x}}_t\right\|^2$$

and the choice of $\gamma \leq \frac{1}{4L}$ makes the term in the bracket positive $(\frac{1}{2})$.

Proof of Lemma 12.7 I

We prove this lemma by recursion.
Note that for any $\beta > 0$: $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\beta)\|\mathbf{a}\|^2 + (1+1/\beta)\|\mathbf{b}\|^2$.

$$
\begin{aligned}
\mathbb{E}R_{t+1} &= \mathbb{E}\left\|\mathbf{x}_{t+1} - \tilde{\mathbf{x}}_{t+1}\right\|^2 \\
&= \mathbb{E}\Big\|\underbrace{\mathbf{x}_t - \tilde{\mathbf{x}}_t}_{=\mathbf{e}_t} + \gamma\mathbf{g}_t - \mathbf{v}_t\Big\|^2 \\
&= \mathbb{E}\left\|\mathbf{e}_t + \gamma\mathbf{g}_t - \mathcal{C}(\mathbf{e}_t + \gamma\mathbf{g}_t)\right\|^2 \\
&\leq (1-\delta)\mathbb{E}\left\|\mathbf{e}_t + \gamma\mathbf{g}_t\right\|^2 \\
&\leq (1-\delta)(1+\beta)\mathbb{E}\mathbb{R}_t + (1-\delta)(1+1/\beta)\gamma^2\mathbb{E}\left\|\mathbf{g}_t\right\|^2 \\
&\leq (1-\delta)(1+\beta)\mathbb{E}\mathbb{R}_t + (1-\delta)(1+1/\beta)\gamma^2 B^2 \\
&\leq \underbrace{(1-\delta/2)\mathbb{E}\mathbb{R}_t}_{\text{exponentially decreasing,}} + \underbrace{\frac{2(1-\delta)\gamma^2}{\delta}B^2}_{\text{residual}} \qquad (*)
\end{aligned}
$$

$\|$

of the form $\|y - \mathcal{C}(y)\|^2$
$\leq (1-\delta)\cdot\|y\|^2$

for the choice $\beta = \frac{\delta}{2(1-\delta)}$ such that $(1+1/\beta) = (2-\delta)/\delta \leq 2/\delta$.

Proof of Lemma 12.7 II

Now we plug-in the bound on $\mathbb{E}R_t$:

$$
\begin{aligned}
ER_{t+1} &\le (1-\delta/2)\left(\frac{4(1-\delta)\gamma^2 B^2}{\delta^2}\right) + \frac{2(1-\delta)\gamma^2 B^2}{\delta} \\
&\le \frac{4(1-\delta)\gamma^2 B^2}{\delta^2}
\end{aligned}
$$

Note that $(1-\delta/2)\frac{2}{\delta^2} + \frac{1}{\delta} = \frac{2}{\delta^2}$.

Discussion

- Only the higher order terms depend on δ.
- "compression for free" with error feedback
- Intuition: gradients stored in the error buffer $\mathbf{e}_t$ are transmitted with a delay τ. Here $\frac{1}{\delta} \approx \tau$ and the results are qualitatively similar.

Extensions:

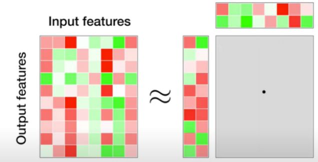
- to multiple workers $n > 1$
- here we assumed $\mathbf{x}_t$ is not compressed. The same feedback-mechanism can be used to compress also the broadcast communication.
- In practice: most relevant are compressors that support efficient aggregation (all-reduce, all-gather).

Outlook

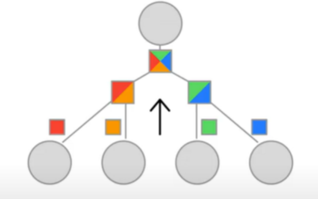
Optimization in Practice

Compression in Practice—PowerSGD [VKJ19]

- Low-rank approximation of weight matrix (power iteration)

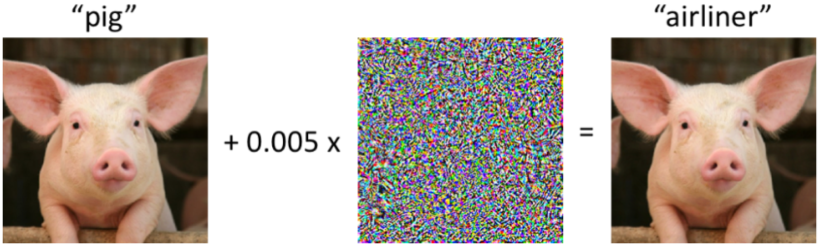


- Efficient all-reduce



- with error-feedback
- Used for large-scale transformer training (DALL-E by OpenAI).

Adversarial Attacks (at inference time)



- Standard training: $\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{a}_i)$ $\qquad$ $\nabla_{\mathbf{x}} f$ change model
- Attacking: $\qquad$ $\nabla_{\mathbf{a}_i} f$ change data

$$\max_{\|\mathbf{a}-\mathbf{a}_i\|\leq\epsilon} f(\mathbf{x}, \mathbf{a})$$

- Algorithm: projected gradient descent $\qquad$ More info here

Other Aspects

- Robustness
  - Byzantine-robust training
- Privacy
  - Secure Multiparty Computation
  - Differential Privacy
  - Privacy/inference Attacks
- machine learning systems
  - decentralized
  - heterogeneous hardware
- Practical tricks
  - limited precision operations
  - number formats for DL
  - feature hashing
- ...

Thanks!

www.sstich.ch

Please reach out if you want to continue working on one of these (or other) topics.
(Master Thesis, HiWi and PhD positions available on a regular basis.)

Bibliography I

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic.
QSGD: Communication-efficient SGD via gradient quantization and encoding.
In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 1709–1720. Curran Associates, Inc., 2017.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi.
Sparsified SGD with memory.
In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 4452–4463. Curran Associates, Inc., 2018.

Sebastian U. Stich and Sai P. Karimireddy.
The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication.
Journal of Machine Learning Research (JMLR), 2020.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi.
PowerSGD: Practical low-rank gradient compression for distributed optimization.
In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 1626–1636. Curran Associates, Inc., 2019.