Optimization for Machine Learning

Lecture 7: Distributed Optimization II

Sebastian Stich

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

May 28, 2024

Group Project

- Work on a research question (related to the course) in a small team!
- Present your result with a **poster** and a short (3 page) **report**.
- A list of project ideas is available here.
  - pick one project (or propose your own)

Hints:

- state the research question/hypothesis clearly!
- only include claims that are supported by evidence you provide!
- the contact person can help you! (but **not** last-minute!)

Group Project Timeline

- Group registration between May 28 – June 4 (register on CMS)
  - groups of 2–3
- before June 18: get in touch with the contact person and schedule a meeting!
  - read the related literature
  - prepare a list of **research goals and tasks**
- before June 25: meet with your contact person
  - zoom meeting, 30-60min, can also be in-person
  - **discuss your research plan**
  - **ask questions** about things you do not understand
- July 16: Poster presentation (& suggested report submission)
  - (note that the **poster printing deadline** is a bit earlier, TBA!)
- July 26: last possible date to submit the report

There will be no exercise sheets in the weeks of June 25/July 2 — you can also discuss the project in the exercise session.

Group Project Grading

- (50%) methodology and execution
  - are claims clearly stated and verified by evidence?
  - is the research question related to the course?
- (50%) presentation of the results
  - poster & poster presentation
  - final report
- bonus points for highly creative questions, interesting results, outstanding presentations, etc.

If the project is not passed, there will be an option to hand in a revised version in August.

Asynchronous SGD

(wrapping up)

Hogwild!

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, stepsize γ, accessible memory location to store $\mathbf{x} \in \mathbb{R}^d$

At iteration t (in parallel):

$$
\begin{aligned}
&\mathbf{x}_t \leftarrow \mathbf{x} && \text{(inconsistent read of the memory } \mathbf{x}\text{)} \\
&\mathbf{g}_t = \mathbf{g}(\mathbf{x}_t) && \text{(stochastic gradient)} \\
&\textbf{for } i \in [d] && \text{(atomic coordinate write)} \\
&\quad [\mathbf{x}]_i := [\mathbf{x}]_i - \gamma[\mathbf{g}_t]_i &&
\end{aligned}
$$

Theorem

Theorem (Lecture-7).1 ([SK20, SMJ21])

Let $f : \mathbb{R}^d \to R$ be L-smooth with $F_0 = f(\mathbf{x}_0) - f^\star$. Then there exists a stepsize $\gamma \leq \gamma_{\text{crit}} := \frac{1}{10L(M+\tau)}$ such that after T steps of delayed SGD (with atomic vector operations):

$$\min_{t \leq T} \mathbb{E} \|\nabla f(\mathbf{x}_t)\|^2 = \mathcal{O}\left(\frac{F_0 L(M+\tau)}{T} + \frac{\sqrt{LF_0\sigma^2}}{\sqrt{T}}\right).$$

- Mini-Batch SGD can be seen as a variant of delayed SGD with $\tau = b$.
- We recover the mini-batch SGD result when considering the same number of gradient computations, i.e. $T \to Tb$ and replacing $\tau \to b$).

Proof I

The main ingredient for the proof is to define a virtual sequence $\tilde{\mathbf{x}}_t$ of iterates, $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$, defined as [MPP+17, SK20]

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

Lemma (Lecture-7).2 (Decrease)

For $\gamma \leq \gamma_{\mathrm{crit}}$ *it holds*

$$\mathbb{E} f(\tilde{\mathbf{x}}_{t+1}) \leq \mathbb{E} f(\tilde{\mathbf{x}}_t) - \frac{\gamma}{4} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma^2 L \sigma^2}{2} + \frac{\gamma L^2}{2} \mathbb{E} \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2$$

Lemma (Lecture-7).3 (Difference)

For $\gamma \leq \gamma_{\mathrm{crit}}$ *it holds* *here* $(t-\tau)_+ = \max\{0, t-\tau\}$

$$\mathbb{E} \|\mathbf{x}_t - \tilde{\mathbf{x}}_t\|^2 \leq \frac{1}{50 L^2 \tau} \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E} \|\nabla f(\mathbf{x}_k)\|^2 + \frac{\gamma}{5L} \sigma^2 .$$

Proof II

Plug (Difference) into (Decrease), re-arrange, and divide by γ:

$$\frac{1}{4}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{1}{\gamma}\left(\mathbb{E}f(\tilde{\mathbf{x}}_t) - \mathbb{E}f(\tilde{\mathbf{x}}_{t+1})\right) + \frac{\gamma L\sigma^2}{2} + \frac{1}{100\tau}\sum_{k=(t-\tau)_+}^{t-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_k)\right\|^2 + \frac{\gamma L\sigma^2}{10}$$

Now we average over T. Note that the highlighted $\left\|\nabla f(\mathbf{x}_k)\right\|^2$ terms appear at most τ times.

$$\frac{1}{4T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{\Delta}{\gamma T} + \gamma L\sigma^2 + \frac{1}{100T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2$$

Note that $\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2$ appears on both sides, with $\frac{1}{4T} - \frac{1}{100T} \geq \frac{1}{5T}$.

$$\frac{1}{5T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{F_0}{\gamma T} + \gamma L\sigma^2\,.$$

Now the result follows by tuning γ (in the same way as before).

Proof of Lemma (Decrease)

This follows our standard path, with one small trick. By L-smoothness (at $\tilde{\mathbf{x}}_t$):

$$\begin{aligned}\mathbb{E}[f(\tilde{\mathbf{x}}_{t+1}) &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \gamma \nabla f(\tilde{\mathbf{x}}_t)^\top \mathbf{g}_t + \tfrac{\gamma^2 L}{2} \mathbb{E}\left\|\mathbf{g}_t\right\|^2 \\ &\leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \gamma \nabla f(\tilde{\mathbf{x}}_t)^\top \nabla f(\mathbf{x}_t) + \frac{\gamma^2 L}{2}\left((M+1)\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \sigma^2\right)\end{aligned}$$

Now:

$$\begin{aligned}-\nabla f(\tilde{\mathbf{x}}_t)^\top \nabla f(\mathbf{x}_t) &= -\left(\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)\right)^\top \nabla f(\mathbf{x}_t) \\ &= -\left\|\nabla f(\mathbf{x}_t)\right\|^2 - \left(\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t)\right)^\top \nabla f(\mathbf{x}_t) \\ &\leq -\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \tfrac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \tfrac{1}{2}\left\|\nabla f(\tilde{\mathbf{x}}_t) - \nabla f(\mathbf{x}_t)\right\|^2 \\ &\leq -\tfrac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \tfrac{L^2}{2}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\right\|^2\end{aligned}$$

where we used $(-\mathbf{a}^\top \mathbf{b}) \leq \frac{1}{2}\left\|\mathbf{a}\right\|^2 + \frac{1}{2}\left\|\mathbf{b}\right\|^2$.

Now with $\gamma \leq \frac{1}{2L(M+\tau)} \leq \frac{1}{2L(M+1)}$:

$$\mathbb{E}[f(\tilde{\mathbf{x}}_{t+1}) \leq \mathbb{E}[f(\tilde{\mathbf{x}}_t)] - \frac{\gamma}{4}\left\|\nabla f(\mathbf{x}_t)\right\|^2 + \frac{\gamma^2 L \sigma^2}{2} + \frac{\gamma L^2}{2}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\right\|^2$$

Proof of Lemma (Difference)

Note that $\mathbf{x}_t = \mathbf{x}_0 - \gamma \sum_{k \in \mathcal{I}_t} \mathbf{g}_k$ and $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \gamma \sum_{k=0}^{t-1} \mathbf{g}_k$ and define $\xi_k = \mathbf{g}_k - \nabla f(\mathbf{x}_k)$, with $\mathbb{E}[\xi_k] = 0$.

Then

$$\mathbb{E}\left\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\right\|^2 = \gamma^2 \mathbb{E}\left\|\sum_{k \in \mathcal{J}_t} \mathbf{g}_k\right\|^2 \leq 2\gamma^2 \mathbb{E}\left\|\sum_{k \in \mathcal{J}_t} \nabla f(\mathbf{x}_k)\right\|^2 + 2\gamma^2 \mathbb{E}\left\|\sum_{k \in \mathcal{J}_t} \xi_k\right\|^2$$

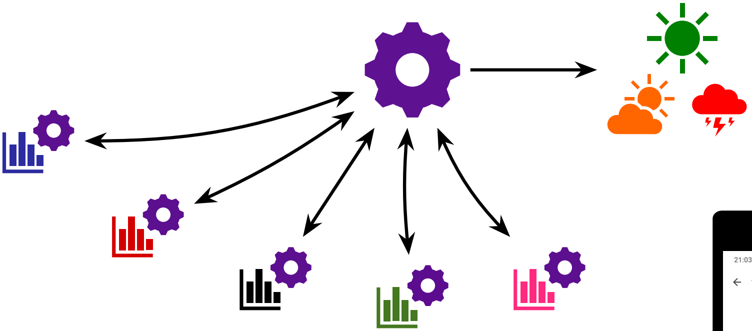
$$\leq 2\gamma^2 \tau \sum_{k=(t-\tau)_+}^{t-1} \left\|\nabla f(\mathbf{x}_k)\right\|^2 + 2\gamma^2 M \sum_{k=(t-\tau)_+}^{t-1} \left\|\nabla f(\mathbf{x}_k)\right\|^2 + 2\gamma^2 \tau \sigma^2$$

Where we used

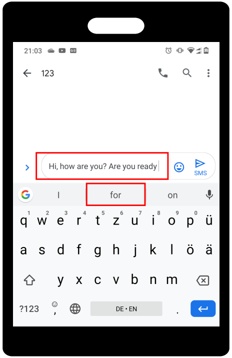
- $\left\|\nabla f(\mathbf{x}_k) + \xi_k\right\|^2 \leq 2\left\|\nabla f(\mathbf{x}_k)\right\|^2 + 2\left\|\xi_k\right\|^2$
- $\left\|\sum_{k=1}^{\tau} \mathbf{a}_k\right\|^2 \leq \tau \sum_{k=1}^{\tau} \left\|\mathbf{a}_k\right\|^2$
- $\mathbb{E}\left\|\sum_{k=1}^{\tau} \xi_k\right\|^2 = \sum_{k=1}^{\tau} \mathbb{E}\left\|\xi_k\right\|^2$ (independent noise)

Federated Learning

Example: Federated Learning [MMR$^+$17, KMea21]



- private data stays on device
- server coordinates training and aggregates focused updates



Training Objective

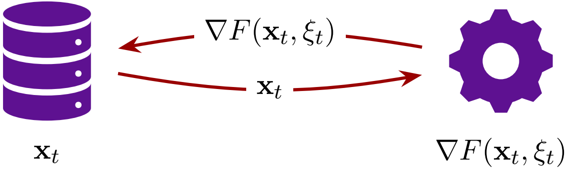
$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \underbrace{f_i(\mathbf{x})}_{\text{data } \mathcal{D}_i \text{ on client } i} \right] \qquad f_i(\mathbf{x}) = \begin{cases} \mathbb{E}_{\xi\sim\mathcal{D}_i} F(\mathbf{x},\xi) \\ \frac{1}{m}\sum_{j=1}^{m} f_{ij}(\mathbf{x})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

- Collaboratively solve **a (joint)** machine learning problem
- **efficiently**, in terms of:
  - computation (stochastic gradients, mini-batches),
  - communication (server $\leftrightarrow$ client).

Other very relevant scenarios:

• personalization • heterogenity • privacy • robustness

Communication Bottleneck



algorithm	rounds	gradients (total)
mini-batch SGD batch size $b = 1$	$\mathcal{O}\left(\frac{\sigma^2}{n\mu\epsilon} + \frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{\mu\epsilon} + \frac{nL}{\mu}\log\frac{1}{\epsilon}\right)$
mini-batch SGD batch size b	$\mathcal{O}\left(\underbrace{\frac{\sigma^2}{nb\mu\epsilon}}_{\ll} + \underbrace{\frac{L}{\mu}\log\frac{1}{\epsilon}}_{=}\right)$	$\mathcal{O}\left(\frac{\sigma^2}{\mu\epsilon} + \frac{nbL}{\mu}\log\frac{1}{\epsilon}\right)$

Local SGD

To parallelize or not to parallelize?

Which algorithm is "better"?

- mini-batch SGD with batch size b and T iterations,
- SGD with bT iterations?

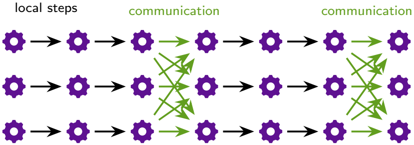
(Both can access the same #oracle calls, $C = bT$). Answer: it depends (on ϵ).

Thought experiment: assume $b \to \infty$ or $T \to \infty$, while keeping C constant.

- Mini-batch SGD ($b \to \infty$) will perform $T = \frac{C}{b} \leq 1$ iteration (and stay at $\mathbf{x}_0$)
- SGD with $b = 1$ will perform $T \to \infty$ steps (potentially converging to $\mathbf{x}^\star$).

Is there a way to "interpolate" between the two extremes?

Local SGD



Notation/Setting

- n machines
- $f_i(\mathbf{x})$ denote the function (data) available locally at node i
- local gradient oracle $\mathbb{E}[\mathbf{g}^{(i)}(\mathbf{x})] = \nabla f_i(\mathbf{x})$, $\forall i \in [n]$, with bounded variance:

$$\mathbb{E}\left\|\mathbf{g}^{(i)}(\mathbf{x}) - \nabla f_i(\mathbf{x})\right\|^2 \leq \sigma^2$$

- let $\mathbf{x}_t^{(i)} \in \mathbb{R}^d$ denote the local iterate at node i, $\forall i \in [n]$

Local SGD

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, $\mathbf{x}_0^{(i)} = \mathbf{x}_0$, $\forall i \in [n]$, stepsize γ, $\tau \geq 1$ (number of local steps)

At iteration t (in parallel on all nodes $i \in [n]$):

$\mathbf{g}_t^i = \mathbf{g}^{(i)}(\mathbf{x}_t^{(i)})$ (stochastic gradient locally on each node)

if $t+1$ is a multiple of τ :

$$\mathbf{x}_{t+1}^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_t^{(i)} - \gamma \mathbf{g}_t^i\right) \qquad \text{(global averaging)}$$

otherwise:

$$\mathbf{x}_{t+1}^{(i)} = \mathbf{x}_t^{(i)} - \gamma \mathbf{g}_t^i \qquad \text{(local step)}$$

Homegeneous Functions

For simplicity, assume

$$f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_n(\mathbf{x})$$

(note that in general $\mathbf{g}_t^i \neq \mathbf{g}_t^j$ for $i \neq j$).

- This means stochastic gradients are uniformly sampled from the whole dataset (similar as for mini-batch SGD).
- For $\tau = 1$ Local SGD is identical to mini-batch SGD with batch size $b = n$.

Theorem (Lecture-7).4 (Homegeneous Case, [Sti19, KLB⁺20])

Let $f_i \colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth, $\forall i \in [n]$ and $f_i = f_j$, $\forall i, j \in [n]$, with $\Delta = f(\mathbf{x}_0) - f^\star$. Then there exists a stepsize $\gamma \leq \gamma_{\text{crit}} := \frac{1}{20L\tau}$ such that after T steps (that is, T/τ communication rounds) of Local SGD it holds

$$\min_{t \leq T} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}_t) \right\|^2 = \mathcal{O}\left(\frac{\Delta L \tau}{T} + \frac{(\Delta L \sigma)^{2/3} \tau^{1/3}}{T^{2/3}} + \frac{\sqrt{L \Delta \sigma^2}}{\sqrt{Tn}} \right),$$

with $\bar{\mathbf{x}}_t := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_t^{(i)}$.

Discussion

- Linear speedup if $\sigma^2 > 0$: the variance decreases linearly in the number of oracle calls (Tn). This is optimal.
- The deterministic optimization term (the term not depending on σ^2) is impacted by τ (similarly as with mini-batch SGD).
  - ideally, we would have hoped to see there $\mathcal{O}\left(\frac{\Delta L}{T}\right)$ (= progress in every iteration) vs. $\mathcal{O}\left(\frac{\Delta L \tau}{T}\right)$ (= progress every communication round)
- The theorem shows almost the same convergence as for mini-batch SGD, up to the higher order $\mathcal{O}(T^{-2/3})$ term.
  - There is no clear winner: see also [WPS+20].

Performance in Practice [LSPJ20]

ResNet-20 on CIFAR-10 (IID data)

	Top-1 acc.	local gradients	communication
Mini-batch SGD ($n = 16$, $\tau = 128$)	92.5%	2048	-
Mini-batch SGD ($n = 16$, $\tau = 1024$)	76.3%	16384	$\div$ 8
Local-SGD ($n = 16$, $\tau = 8 \times 128$)	92.0%	16384	$\div$ 8

Proof I

We will again use the virtual sequence technique. As virtual sequence we consider $\bar{\mathbf{x}}_t$ (note that the average is not computed in every iteration).

Lemma (Lecture-7).5 (Decrease)

For $\gamma \leq \frac{1}{4L}$ it holds

$$\mathbb{E}f(\bar{\mathbf{x}}_{t+1}) \leq Ef(\bar{\mathbf{x}}_t) - \frac{\gamma}{4}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 + \gamma^2 L\frac{\sigma^2}{n} + \frac{\gamma L^2}{n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_t^{(i)} - \bar{\mathbf{x}}_t\right\|^2$$

Lemma (Lecture-7).6 (Difference)

For $\gamma \leq \gamma_{\text{crit}} = \frac{1}{20L\tau}$, with the notation for $R_t = \frac{1}{n}\sum_{i=1}^n \left\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(i)}\right\|^2$, it holds

$$\mathbb{E}R_t \leq \frac{1}{20L^2\tau}\sum_{j=(t-1)-k}^{t-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_j)\right\|^2 + 5\gamma^2\tau\sigma^2$$

where $(t-1)-k$ denotes the index of the last communication round ($k \leq \tau - 1$).

Proof II

Plug (Difference) into (Decrease), re-arrange and divide by γ:

$$\frac{1}{4}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{1}{\gamma}\left(\mathbb{E}f(\bar{\mathbf{x}}_t) - \mathbb{E}f(\bar{\mathbf{x}}_{t+1})\right) + \gamma L\frac{\sigma^2}{n} + \frac{1}{20\tau}\sum_{j=(t-1)-k}^{t-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_j)\right\|^2 + 5\gamma^2 L^2 \tau\sigma^2$$

Now we divide by T and sum over $t = 0, \dots, T-1$:

$$\frac{1}{4T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{1}{\gamma T}\sum_{t=0}^{T-1}\left[\left(\mathbb{E}f(\bar{\mathbf{x}}_t) - \mathbb{E}f(\bar{\mathbf{x}}_{t+1})\right) + \frac{1}{20T}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2\right] + \gamma^2 L\frac{\sigma^2}{n} + 5\gamma L^2\tau\sigma^2$$

Note that $\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2$ appears on both sides, with $\frac{1}{4T} - \frac{1}{20T} = \frac{1}{5T}$.

$$\frac{1}{5T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{\Delta}{\gamma T} + \gamma L\frac{\sigma^2}{n} + 5\gamma^2 L^2\tau\sigma^2 \,.$$

Now the result follows by tuning γ (see Exercise Sheet 7).

Proof of Lemma (Decrease)

By L-smoothness:

$$\mathbb{E} f(\bar{\mathbf{x}}_{t+1}) \leq \mathbb{E} f(\bar{\mathbf{x}}_t) - \frac{\gamma}{n} \sum_{i=1}^n \nabla f(\bar{\mathbf{x}}_t)^\top \mathbf{g}_t^i + \frac{\gamma^2 L}{2} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{g}_t^i \right\|^2$$

$$\leq \mathbb{E} f(\bar{\mathbf{x}}_t) - \frac{\gamma}{n} \sum_{i=1}^n \nabla f(\bar{\mathbf{x}}_t)^\top \nabla f_i(\mathbf{x}_t^{(i)}) + \frac{\gamma^2 L \sigma^2}{2n} + \frac{\gamma^2 L}{2} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_t^{(i)}) \right\|^2$$

Similarly as we have seen before, by adding and subtracting $\nabla f(\bar{\mathbf{x}}_t)$:

$$-\frac{1}{n} \sum_{i=1}^n \nabla f(\bar{\mathbf{x}}_t)^\top \nabla f_i(\mathbf{x}_t^{(i)}) = -\nabla f(\bar{\mathbf{x}}_t)^\top \nabla f(\bar{\mathbf{x}}_t) + \frac{1}{n} \sum_{i=1}^n \nabla f(\bar{\mathbf{x}}_t)^\top \left(\nabla f(\bar{\mathbf{x}}_t) - \nabla f_i(\mathbf{x}_t^{(i)}) \right)$$

$$\leq -\frac{1}{2} \left\| \nabla f(\bar{\mathbf{x}}_t) \right\|^2 + \frac{1}{2} \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_t^{(i)}) \right\|^2$$

$$\leq -\frac{1}{2} \left\| \nabla f(\bar{\mathbf{x}}_t) \right\|^2 + \frac{L^2}{2n} \sum_{i=1}^n \left\| \bar{\mathbf{x}}_t - \mathbf{x}_t^{(i)} \right\|^2$$

Note: $-\mathbf{a}^\top \mathbf{b} \leq \frac{1}{2} \left\| \mathbf{a} \right\|^2 + \frac{1}{2} \left\| \mathbf{b} \right\|^2$.

Continued

And by adding and subtracting $\nabla f(\bar{\mathbf{x}}_t)$ in the last term: $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$

$$\begin{aligned}
\frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2 + \|\nabla f(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \frac{L^2}{n}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(i)}\right\|^2 + \|\nabla f(\bar{\mathbf{x}}_t)\|^2
\end{aligned}$$

Now we plug everything together, and use $\gamma \leq \frac{1}{4L}$.

$$\mathbb{E}f(\bar{\mathbf{x}}_{t+1}) \leq \mathbb{E}f(\bar{\mathbf{x}}_t) + \left(\gamma^2 L - \frac{\gamma}{2}\right)\|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \frac{\gamma^2 L\sigma^2}{2n} + \left(\frac{\gamma L^2}{2} + \gamma^2 L^3\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_t - \mathbf{x}_t^{(i)}\right\|^2$$

Proof of Lemma (Difference)

Note that if t is a multiple of τ, then $R_t = 0$ and there is nothing to prove. Otherwise note that

$$\begin{aligned}\mathbb{E}R_{t+1} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}^{i}\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}_{t} - \mathbf{x}_{t}^{i} + \gamma\mathbf{g}_t^i - \gamma\bar{\mathbf{g}}_t\right\|^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}_{t} - \mathbf{x}_{t}^{i} + \gamma\nabla f_i(\mathbf{x}_t^i) - \gamma\bar{\mathbf{v}}_t\right\|^2 + \gamma^2\sigma^2\,,\end{aligned}$$

where $\bar{\mathbf{g}}_t := \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_t^i$ and $\bar{\mathbf{v}}_t := \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t^{(i)})$ denote the average of the client gradients. With the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\tau^{-1})\|\mathbf{a}\|^2 + 2\tau\|\mathbf{b}\|^2$ for $\tau \geq 1$, $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$, we continue:

$$\begin{aligned}\mathbb{E}R_{t+1} &\leq \left(1+\frac{1}{\tau}\right)\mathbb{E}R_t + \frac{2\tau\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f(\mathbf{x}_t^{(i)}) - \bar{\mathbf{v}}_t\right\|^2 + \gamma^2\sigma^2 \\ &\leq \left(1+\frac{1}{\tau}\right)\mathbb{E}R_t + \frac{2\tau\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 + \gamma^2\sigma^2 \\ &\leq \left(1+\frac{1}{\tau}\right)\mathbb{E}R_t + \frac{2\tau\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}\left(2\left\|\nabla f_i(\bar{\mathbf{x}}_t)\right\|^2 + 2\left\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\bar{\mathbf{x}}_t)\right\|^2\right) + \gamma^2\sigma^2\end{aligned}$$

$(*)$

Continued

We now use $f_1 = f_2 = \cdots = f_n$ and $\gamma \leq \frac{1}{20L\tau}$ to simplify:

$$\begin{aligned}
\mathbb{E}R_{t+1} &\leq \left(1 + \frac{1}{\tau}\right)\mathbb{E}R_t + \frac{1}{100L^2\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 + \frac{1}{100\tau}\mathbb{E}R_t + \gamma^2\sigma^2 \\
&\leq \left(1 + \frac{3}{2\tau}\right)\mathbb{E}R_t + \frac{1}{100L^2\tau}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 + \gamma^2\sigma^2
\end{aligned}$$

The lemma now follows by unrolling, and noting that $\left(1 + \frac{3}{2\tau}\right)^j \leq 5$ for all $0 \leq j \leq \tau$.

Lecture 7 Recap

- Federated Learning
  - studied the convergence properties of local SGD
  - in practice: FedAvg

- Homogeneous/IID optimization setting
  - might not be realistic for real-world applications!

Bibliography I

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich.
A unified theory of decentralized SGD with changing topology and local updates.
In *37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

Peter Kairouz, H. Brendan McMahan, and et al.
Advances and open problems in federated learning.
Foundations and TrendsÂ® in Machine Learning, 14(1–2):1–210, 2021.

Tao Lin, Sebastian U. Stich, Kumar K. Patel, and Martin Jaggi.
Don't use large mini-batches, use local SGD.
International Conference on Learning Representations (ICLR), 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
Communication-efficient learning of deep networks from decentralized data.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

Bibliography II

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan.
Perturbed iterate analysis for asynchronous stochastic optimization.
SIAM Journal on Optimization, 27(4):2202–2229, 2017.

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.

Sebastian Stich, Amirkeivan Mohtashami, and Martin Jaggi.
Critical parameters for scalable distributed learning with large batches and asynchronous updates.
In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4042–4050. PMLR, 13–15 Apr 2021.

Bibliography III

Sebastian U. Stich.
Local SGD converges fast and communicates little.
International Conference on Learning Representations (ICLR), 2019.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro.
Is local SGD better than minibatch SGD?
In *37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

b

Discussion

Discussion

Discussion