Labs
Optimization for Machine Learning
Spring 2024

Saarland University
CISPA Helmholtz Center for Information Security
Sebastian Stich
TAs: Yuan Gao & Xiaowen Jiang
https://cms.cispa.saarland/optml24/

Problem Set 7 — Solutions (Local SGD)

1 Local SGD on Heterogeneous Functions

Consider the (generalized) example from the lecture, with $f\colon \mathbb{R} \to \mathbb{R}$ defined as:

$$f_1(x) = \frac{1}{2}x^2 \qquad f_2(x) = a(x-1)^2 \qquad f(x) = \frac{1}{2}\left(f_1(x) + f_2(x)\right),$$

for $a \geq 0$. Verify that the optimal solution $x^\star := \operatorname{argmin} f(x)$ is given as $x^\star = \frac{2a}{1+2a}$.

1.1 The optimal solution is not a fix point of Local SGD

Consider local SGD with stepsize $\gamma > 0$, and $\tau = 2$ local steps. Prove that when we start local SGD at $x_0 = x^\star$ we end up at

$$x_2 = x^\star + \frac{(a-2a^2)\gamma^2}{1+2a}$$

after the first averaging round.

1.2 Similarity

Based the previous observation, can you derive conditions under which $x_2 = x^\star$, i.e. the optimal solution is a fixed point? Do these conditions also hold for $\tau > 2$ local steps?

Proof. By setting the gradient to zero, $0 = \nabla f_1(x^\star) + \nabla f_2(x^\star) = x^\star + 2a(x^\star - 1)$ we deduce $x^\star = \frac{2a}{1+2a}$. Compute first the local iterates after two steps of local SGD, but before averaging. In the lecture we denoted these iterates as $x_2^{(i)'}$, for $i \in \{1,2\}$. We obtain:

$$x_2^{(1)'} = (1-\gamma)^2 x_0 = (1-\gamma)^2 \frac{2a}{1+2a}$$
$$x_2^{(2)'} = 1 + (1-a\gamma)^2(x_0 - 1) = 1 - \frac{(1-a\gamma)^2}{1+2a}$$

and finally

$$x_2 = \frac{1}{2}\left(x_2^{(1)'} + x_2^{(2)'}\right) = \frac{2a}{1+2a} + \frac{(a-2a^2)\gamma^2}{1+2a} = x^\star\left(1 + \frac{(1-2a)\gamma^2}{2}\right).$$

From this we see that when $a = \frac{1}{2}$, then $x^\star$ is a fix point. $\square$

2 Verify the proof of the Local SGD Theorem (general case):

In the lecture, we left out two steps:

- Plugging the (Difference) lemma into the (Decrease) lemma, and rearranging the terms to obtain

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

- And the tuning of the stepsize (with respect to the constraint $\gamma \leq \frac{1}{10L\tau}$).

Proof. Recall the (Decrease) lemma

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

and the (Difference) lemma

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}||\mathbf{x}_t^{(i)} - \bar{\mathbf{x}}_t||^2] \leq \frac{1}{10L^2\tau}\sum_{j=(t-1)-k}^{t-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_j)||^2] + 5\gamma^2\sigma^2\tau + 40\gamma^2\tau^2\zeta^2$$

Plug (Difference) into (Decrease), rearrange and divide by γ

$$\frac{1}{4}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_t)||^2] \leq \frac{1}{\gamma}(\mathbb{E}[f(\bar{\mathbf{x}}_t)] - \mathbb{E}[f(\bar{\mathbf{x}}_{t+1})]) + \gamma L\frac{\sigma^2}{n} + \frac{1}{10\tau}\sum_{j=(t-1)-k}^{t-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_j)||^2] + 5\gamma^2 L^2\sigma^2\tau + 40\gamma^2 L^2\tau^2\zeta^2$$

Divide by T and sum over $t = 0, ..., T-1$

$$\frac{1}{4T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_t)||^2] \leq \frac{f(\mathbf{x}_0) - f^\star}{\gamma T} + \frac{1}{10T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_t)||^2] + \gamma L\frac{\sigma^2}{n} + 5\gamma^2 L^2\sigma^2\tau + 40\gamma^2 L^2\tau^2\zeta^2$$

Use $\frac{1}{4T} - \frac{1}{10T} \geq \frac{1}{8T}$ and rearrange

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_t)||^2] &\leq 8\frac{f(\mathbf{x}_0) - f^\star}{\gamma T} + 8\gamma L\frac{\sigma^2}{n} + 40\gamma^2 L^2\sigma^2\tau + 320\gamma^2 L^2\tau^2\zeta^2 \\ &= \mathcal{O}\left(\frac{\Delta}{\gamma T} + \gamma L\frac{\sigma^2}{n} + \gamma^2 L^2(\tau^2\zeta^2 + \tau\sigma^2)\right)\end{aligned}$$

Use Exercise 8.1 with $A = \Delta$, $B = \frac{L\sigma^2}{n}$, $C = L^2(\tau^2\zeta^2 + \tau\sigma^2)$ and $D = L\tau$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\nabla f(\bar{\mathbf{x}}_t)||^2] \leq \mathcal{O}\left(\frac{\Delta L\tau}{T} + \left(\frac{L\Delta(\tau\zeta + \sqrt{\tau}\sigma)}{T}\right)^{\frac{2}{3}} + \frac{\sqrt{L\Delta\sigma^2}}{\sqrt{Tn}}\right)$$

□