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 6 — Solutions (Mini-batch and Async)

1 Tuning the Stepsize

Let $A, B, C \geq 0$ and $D > 0$ be given parameters. Consider the expression

$$\Psi(T, \gamma) := \frac{A}{\gamma T} + B\gamma + C\gamma^2$$

depending on T and γ. Show that for any $T \geq 1$

$$\min_{\gamma \leq \frac{1}{D}} \Psi(T, \gamma) \leq 2\left(\frac{AB}{T}\right)^{1/2} + 2C^{1/3}\left(\frac{A}{T}\right)^{2/3} + \frac{AD}{T}\,.$$

Hint: Prove the result first for the special case $C = 0$.

Proof. Choosing $\gamma = \min\left\{\left(\frac{A}{BT}\right)^{\frac{1}{2}}, \left(\frac{A}{CT}\right)^{\frac{1}{3}}, \frac{1}{D}\right\} \leq \frac{1}{D}$ we have three cases

- $\gamma = \frac{1}{D}$ and is smaller than both $\left(\frac{A}{BT}\right)^{\frac{1}{2}}$ and $\left(\frac{A}{CT}\right)^{\frac{1}{3}}$, then

$$\Psi(T, \gamma) \leq \frac{AD}{T} + \frac{B}{D} + \frac{C}{D^2} \leq \left(\frac{BA}{T}\right)^{\frac{1}{2}} + \frac{DA}{T} + C^{1/3}\left(\frac{A}{T}\right)^{\frac{2}{3}}$$

- $\gamma = \left(\frac{A}{BT}\right)^{\frac{1}{2}} < \left(\frac{A}{CT}\right)^{\frac{1}{3}}$, then

$$\Psi(T, \gamma) \leq 2\left(\frac{AB}{T}\right)^{\frac{1}{2}} + C\left(\frac{A}{BT}\right) \leq 2\left(\frac{AB}{T}\right)^{\frac{1}{2}} + C^{\frac{1}{3}}\left(\frac{A}{T}\right)^{\frac{2}{3}},$$

- The last case, $\gamma = \left(\frac{A}{CT}\right)^{\frac{1}{3}} < \left(\frac{A}{BT}\right)^{\frac{1}{2}}$

$$\Psi(T, \gamma) \leq 2C^{\frac{1}{3}}\left(\frac{A}{T}\right)^{\frac{2}{3}} + B\left(\frac{A}{CT}\right)^{\frac{1}{3}} \leq 2C^{\frac{1}{3}}\left(\frac{A}{T}\right)^{\frac{2}{3}} + \left(\frac{AB}{T}\right)^{\frac{1}{2}}\,. \qquad \square$$

2 Bias-Variance Decomposition

Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a differentiable function and $\mathbf{g}(\mathbf{x})$ a gradient oracle $\mathbf{g} \colon \mathbb{R}^d \to \mathbb{R}^d$ with $\mathbb{E}[\mathbf{g}(\mathbf{x})] = \nabla f(\mathbf{x})$, $\mathbb{E}\left\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq M\left\|\nabla f(\mathbf{x})\right\|^2 + \sigma^2$, $\forall \mathbf{x} \in \mathbb{R}^d$. Show that

$$\mathbb{E}\left\|\mathbf{g}(\mathbf{x})\right\|^2 \leq (M+1)\left\|\nabla f(\mathbf{x})\right\|^2 + \sigma^2\,.$$

Proof. We prove that for a random variable X it holds $\mathbb{E}\left\|X - \mathbb{E}[X]\right\|^2 = \mathbb{E}\left\|X\right\|^2 - \left\|\mathbb{E}[X]\right\|^2$. Note that

$$\mathbb{E}\left\|X - \mathbb{E}[X]\right\|^2 = \mathbb{E}\left\|X\right\|^2 - 2\underbrace{\mathbb{E}[X]^\top E[X]}_{=\left\|\mathbb{E}[X]\right\|^2} + \left\|\mathbb{E}[X]\right\|^2 = \mathbb{E}\left\|X\right\|^2 - \left\|\mathbb{E}[X]\right\|^2\,.$$

The statement of the exercise question follows by setting $X = \mathbf{g}(\mathbf{x})$, $\mathbb{E}[X] = \nabla f(\mathbf{x})$. $\square$

3 Hogwild!

Consider the Howild! algorithm from the lecture. We want to prove its convergence under atomic coordinate-writes (in contrasts to atomic vector-writes as studied in the lecture).

3.1 Notation

Suppose we want to express the iterates of the algorithm as

$$\mathbf{x}_t = \mathbf{x}_0 - \gamma \sum_{k=0}^{t-1} \mathbf{J}_k^t \mathbf{g}_k$$

for matrices $\mathbf{J}_k^t \in \mathbb{R}^{d\times d}$, $k < t$. Define $\mathbf{J}_k^t$.
Hint: Considering diagonal matrices suffices.

Proof. We can consider

$$(\mathbf{J}_k^t)_{vv} = \begin{cases} 1 & \text{if } [\mathbf{g}_k]_v \text{ written before } [\mathbf{x}_t]_v \text{ was read,} \\ 0 & \text{otherwise.} \end{cases}$$

□

3.2 "Difference" Lemma

Prove that the difference Lemma still holds (under the same assumptions on f and γ_{crit} as in the lecture):

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

Proof. First, we observe that by definition of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_t$ and the maximal overlap τ, we can write

$$\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2 := \left\|\gamma \sum_{k<t} (\mathbf{J}_k^t - \mathbf{I}_d)\mathbf{g}_k\right\|^2 = \left\|\gamma \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\mathbf{g}_k\right\|^2 ,$$

where $\mathbf{g}_k := \nabla f(\mathbf{x}_k) + \boldsymbol{\xi}_k$ for zero-mean noise terms. Therefore

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}_t\right\|^2 &\overset{①}{\leq} 2\gamma^2 \left(\mathbb{E}\left\| \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\right\|^2 + \mathbb{E}\left\| \sum_{k=(t-\tau)_+}^{t-1} (\mathbf{J}_k^t - \mathbf{I}_d)\boldsymbol{\xi}_k\right\|^2 \right) \\
&\overset{②}{\leq} 2\gamma^2 \left(\tau \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\left\|(\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\right\|^2 + \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\left\|(\mathbf{J}_k^t - \mathbf{I}_d)\boldsymbol{\xi}_k\right\|^2 \right) \\
&\overset{③}{\leq} 2\gamma^2 \left(\tau \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_k)\right\|^2 + \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\left\|\boldsymbol{\xi}_k\right\|^2 \right) \\
&\overset{④}{\leq} 2\gamma^2 \left((\tau + M) \sum_{k=(t-\tau)_+}^{t-1} \mathbb{E}\left\|\nabla f(\mathbf{x}_k)\right\|^2 + \tau\sigma^2 \right) ,
\end{aligned}$$

where we used ① $\left\|\mathbf{a}+\mathbf{b}\right\|^2 \leq 2\left\|\mathbf{a}\right\|^2 + 2\left\|\mathbf{b}\right\|^2$, ② $\left\|\sum_{i=1}^{\tau} \mathbf{a}_i\right\|^2 \leq \tau \sum_{i=1}^{\tau}\left\|\mathbf{a}_i\right\|^2$, and $\mathbb{E}\left\|\sum_{i=1}^{\tau}\boldsymbol{\xi}_i\right\|^2 = \sum_{i=1}^{\tau}\mathbb{E}\left\|\boldsymbol{\xi}_k\right\|^2$, ③ $\left\|(\mathbf{J}_k^t - \mathbf{I}_d)\nabla f(\mathbf{x}_k)\right\|^2 \leq \left\|\mathbf{J}_k^t - \mathbf{I}_d\right\|^2 \left\|\mathbf{g}_k\right\|^2 \leq \left\|\nabla f(\mathbf{x}_k)\right\|^2$, ④ $\mathbb{E}\left\|\boldsymbol{\xi}_k\right\|^2 \leq M\left\|\nabla f(\mathbf{x}_k)\right\|^2 + \sigma^2$. □