L7 - 10/11/23 - Optimization

▷ **Assumptions:**

(1) $|u^T \mathcal{H}f(x)u| \leq L\|u\|^2 \quad \forall x$ (bound on the second deriv.)

(2) $\underbrace{\mathbb{E}[\|\nabla f_{i(k)}(x) - \mathbb{E}[\nabla f_{i(k)}(x)]\|^2]}_{= Var[\nabla f_{i(k)}(x)]} \leq \sigma^2 \quad \forall x$

$\iff$ is about $Var[\nabla f_{i(k)}(x)] \leq \sigma^2$

$\iff \mathbb{E}[\|\nabla f_{i(k)}(x) - \nabla f(x)\|^2] \leq \sigma^2$

$$\mathbb{E}[\|\nabla f_{i(k)}(x) - \nabla f(x)\|^2]$$

Expectation is linear $\rightsquigarrow$

$$= \mathbb{E}[\|\nabla f_{i(k)}(x)\|^2] - 2\mathbb{E}[<\nabla f_{i(k)}(x), \nabla f(x)>] + \mathbb{E}[\|\nabla f(x)\|^2]$$

$$= \mathbb{E}[\|\nabla f_{i(k)}(x)\|^2] \underbrace{- 2<\mathbb{E}[\underbrace{\nabla f_{i(k)}(x)}_{= \nabla f(x^{(k)})}], \nabla f(x)>}_{= -2\|\nabla f(x)\|^2} + \underbrace{\mathbb{E}[\|\nabla f(x)\|^2]}_{= \|\nabla f(x)\|^2}$$

$$= \mathbb{E}[\|\nabla f_{i(k)}(x)\|^2] - \|\nabla f(x)\|^2 \leq \sigma^2$$

In conclusion, the bound on variance gives

$$\mathbb{E}[\|\nabla f_{i(k)}(x)\|^2] \leq \sigma^2 + \|\nabla f(x)\|^2$$

By Taylor, $\exists \xi \in \mathbb{R}^n$ s.t.

$$f(x^{(k+1)}) \overset{\uparrow}{=} f(x^{(k)} - a^{(k)}\nabla f_{i(k)}(x^{(k)}))$$

$$= f(x^{(k)}) - a^{(k)} < \nabla f_{i(k)}(x^{(k)}), \nabla f(x^{(k)}) >$$

$$+ \frac{1}{2}(a^{(k)}\nabla f_{i(k)}(x^{(k)}))^T \mathcal{H}f(\xi)(a^{(k)}\nabla f_{i(k)}(x^{(k)}))$$

Ass. (1)

$$\Rightarrow f(x^{(k+1)}) \leq f(x^{(k)}) - a^{(k)} < \nabla f_{i(k)}(x^{(k)}), \nabla f(x^{(k)}) >$$

$$+ \frac{L}{2}(a^{(k)})^2 \|\nabla f_{i(k)}(x^{(k)})\|^2$$

$$\mathbb{E}[f(x^{(k+1)})] \leq \mathbb{E}[f(x^{(k)})] - a^{(k)}\|\nabla f(x^{(k)})\|^2$$

$$+ \frac{(a^{(k)})^2 L}{2}(\sigma^2 + \|\nabla f(x^{(k)})\|^2)$$

$$= \mathbb{E}[f(x^{(k)})] - \frac{a^{(k)}}{2}\underbrace{(2 - a^{(k)}L)}_{\geq 1}\|\nabla f(x^{(k)})\|^2$$

$$+ \frac{(a^{(k)})^2 L \cdot \sigma^2}{2}$$

if $a^{(k)}$ is sufficiently small ($a^{(k)} \leq \frac{1}{L}$) ->

$$\le \mathbb{E}[f(x^{(k)})] - \frac{a^{(k)}}{2}\|\nabla f(x^{(k)})\|^2 + \frac{(a^{(k)})^2 L \cdot \sigma^2}{2}$$

$$\Rightarrow a^{(k)}\|\nabla f(x^{(k)})\|^2 \le (\mathbb{E}[f(x^{(k)})] - \mathbb{E}[f(x^{(k+1)})]) + (a^{(k)})^2 L \cdot \sigma^2$$

$$\underset{\substack{\uparrow \\ \text{summing up} \\ \text{order } T-1 \\ \text{steps}}}{\Rightarrow} \sum_{k=0}^{T-1} a^{(k)}\|\nabla f(x^{(k)})\|^2 \le 2(\underbrace{\mathbb{E}[f(x^{(0)})]}_{f(x^{(0)})} - \mathbb{E}[f(x^{(T)})]) + L\sigma^2 \sum (a^{(k)})^2$$

$f(x^{(T)}) \ge f^*$ is the local minimum

$\Rightarrow \mathbb{E}[f(x^{(T)})] \ge f^*$

Hence:

$$\sum_{k=0}^{T-1} a^{(k)}\|\nabla f(x^{(k)})\|^2 \le 2(f(x^{(0)}) - f^*) + L\sigma \sum_{k=0}^{T-1} (a^{(k)})^2$$

Now assume that we do random number t of steps between $0 \dots T-1$ with probability

$$\mathbb{P}(\tau = t) = \frac{a^{(t)}}{\sum_{i=0}^{T-1} a^{(k)}}$$

0	1	2	3		$T-1$
$\frac{a^{(0)}}{\sum a^{(i)}}$	$\frac{a^{(1)}}{\sum a^{(i)}}$	$\frac{a^{(1)}}{\sum a^{(i)}}$			$\frac{a^{(T-1)}}{\sum a^{(i)}}$

$$\mathbb{E}[\|\nabla f(\underset{\substack{\uparrow \\ \text{new randomness}}}{x^{(\tau)}})\|^2] \overset{\text{def}}{=} \sum_{i=0}^{T-1} \|\nabla f(x^{(i)})\|^2 \cdot P(\tau = i)$$

$$= \sum_{i=0}^{T-1} \frac{a^{(i)}\|\nabla f(x^{(i)})\|^2}{\sum_{i=0}^{T-1} a^{(i)}} = \left(\sum_{i=0}^{T-1} a^{(i)}\right)^{-1} \sum_{i=0}^{T-1} a^{(i)}\|\nabla f(x^{(i)})\|^2$$

$$\le \frac{2(f(x^{(0)}) - f^*)}{\sum_{i=0}^{T-1} a^{(i)}} + L\sigma^2 \frac{\sum_{i=0}^{T-1} (a^{(i)})^2}{\sum_{i=0}^{T-1} a^{(i)}}$$

if $a^{(i)} = a \ \forall i$, then $\mathbb{E}[\|\nabla f(x^{(\tau)})\|^2] \le \underset{\substack{\downarrow\, T\to\infty \\ 0}}{\frac{2(f(x^{(0)}) - f^*)}{aT}} + \overbrace{\underset{\substack{\downarrow \\ 0}}{L\sigma^2 a}}^{\text{Noise ball}}$

How to assure convergance (inexpectation)

- $\sum_{i=0}^{T-1} a^{(i)} \xrightarrow[T\to\infty]{} \infty$
- $\sum_{i=0}^{T-1} a^{(i)}$ goes to ∞ faster than $\sum_{i=0}^{T-1} (a^{(i)})^2$

→

For example: $\alpha^{(k)} = \frac{1}{\sqrt{k+1}}$ does the job!

$$\sum_{k=0}^{T-1} \frac{1}{\sqrt{k+1}} \sim \int_0^T \frac{1}{\sqrt{x}}\,dx = 2\sqrt{T}$$

$$\sum_{k=0}^{T-1} \frac{1}{k+1} \sim \int_0^T \frac{1}{x}\,dx = \log T$$

$\sqrt{T}$ grows much faster than $\log T$.

Remark: If f is strictly convex then SGD performs much better: it converges (in expectation) even with constraint learning rate.