

## Lecture 12: 5 October 2023

Instructor: Shashank Vatedka

Scribe: Gautam Singh

**Disclaimer:** These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.

## 12.1 Distributed Stochastic Gradient Descent

Note that if  $\mathbf{h}_1(\mathbf{x})$  and  $\mathbf{h}_2(\mathbf{x})$  are independent stochastic gradients, then  $\frac{1}{2}(\mathbf{h}_1(\mathbf{x}) + \mathbf{h}_2(\mathbf{x}))$  is also a stochastic gradient.

**Definition 12.1.** The **variance** of a stochastic gradient  $\mathbf{h}(\mathbf{x})$  is defined as

$$\text{Var}(\mathbf{h}(\mathbf{x})) \triangleq \mathbb{E} \left[ \|\mathbf{h}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \right]. \quad (12.1)$$

Suppose that  $\mathbf{h}_i$ ,  $1 \leq i \leq k$  are iid stochastic gradients and  $\text{Var}(\mathbf{h}_i) \leq \sigma^2$ . Then,

$$\bar{\mathbf{h}} \triangleq \frac{1}{k} \sum_{i=1}^k \mathbf{h}_i \quad (12.2)$$

is also a stochastic gradient where  $\text{Var}(\bar{\mathbf{h}}) \leq \frac{\sigma^2}{k}$ . We can reduce communication costs by quantizing the stochastic gradients. If an unbiased quantizer is used, then the quantized gradients will also be stochastic.

**Theorem 12.2** (Averaged SGD). Suppose that  $\mathcal{X} \subset \mathbb{R}^d$  is a convex set and  $f : \mathcal{X} \rightarrow \mathbb{R}$  is a convex  $L$ -smooth function, where for some  $L > 0$  and  $\forall x, y$ ,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2. \quad (12.3)$$

Consider an SGD with initial point  $\mathbf{x}_0$ . Then, let

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{x}_0\|_2 \leq R \quad (12.4)$$

and let  $T$  be the number of iterations in the SGD, with learning rate

$$\eta_t = \frac{1}{L + \frac{1}{\gamma}}, \quad \gamma = \frac{R}{\sigma} \sqrt{\frac{2}{T}} \quad (12.5)$$

where  $\sigma$  is the variance of the stochastic gradient. Suppose the SGD generates points  $\mathbf{x}_i$ ,  $1 \leq i \leq T$ . Then,

$$\mathbb{E} \left[ f \left( \frac{1}{T} \sum_{i=1}^T \mathbf{x}_i \right) \right] - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \leq R \sqrt{\frac{2\sigma^2}{T}} + \frac{LR^2}{T} \quad (12.6)$$

Notice that for large  $T$  and  $\sigma^2 = 0$ , the averaged SGD converges to the true minimum.

Note also that the speed of SGD depends on

1. Time to compute *unquantized* stochastic gradients  $\mathbf{h}_i$ .
2. Time complexity of *quantization* for the gradients.
3. Number of GPUs used and resources available.
4. Total communication time.

In these settings, the preferred quantization method is  $k$ -bit randomized rounding, since it is unbiased, and also

$$\mathbb{E} [\|Q(\mathbf{x})\|_0] \leq 2^k (2^k + \sqrt{d}). \quad (12.7)$$

That is, the quantized  $\mathbf{x}$  is sparse. Hence, we can send the values and locations. The total number of bits needed is thus

$$B \leq k\sqrt{d} + \log \binom{d}{k\sqrt{d}} \leq k\sqrt{d} + \mathcal{O}(\sqrt{d} \log d) \leq \mathcal{O}(\sqrt{d} \log d) \quad (12.8)$$