EE6367: Topics in Data Storage and Communications **2023**

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 graidents, then $\frac{1}{2}\left(\mathbf{h_1}(\mathbf{x}) + \mathbf{h_2}(\mathbf{x})\right)$ is also a stochastic gradient.

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

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

Suppose that $\mathbf{h_i},\ 1 \leqslant i \leqslant k$ are iid stochastic gradients and $\operatorname{Var}\left(\mathbf{h_i}\right) \leqslant \sigma^2$. Then,

$$\bar{\mathbf{h}} \triangleq \frac{1}{k}\sum_{i=1}^{k} \mathbf{h_i} \tag{12.2}$$

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

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

$$\left\|\boldsymbol{\nabla} f(\mathbf{x}) - \boldsymbol{\nabla} f(\mathbf{y})\right\|_2 \leqslant L\left\|\mathbf{x} - \mathbf{y}\right\|_2. \tag{12.3}$$

Consider an SGD with initial point $\mathbf{x_0}$. Then, let

$$\sup_{\mathbf{x}\in\mathcal{X}} \left\|\mathbf{x} - \mathbf{x_0}\right\|_2 \leqslant R \tag{12.4}$$

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

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

where σ is the variance of the stochastic gradient. Suppose the SGD generates points $\mathbf{x_i},\ 1 \leqslant i \leqslant T$. Then,

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

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

$$\mathbb{E}\left[\|Q(\mathbf{x})\|_0\right] \leqslant 2^k \left(2^k + \sqrt{d}\right). \tag{12.7}$$

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

$$B \leqslant k\sqrt{d} + \log\binom{d}{k\sqrt{d}} \leqslant k\sqrt{d} + \mathcal{O}\left(\sqrt{d}\log d\right) \leqslant \mathcal{O}\left(\sqrt{d}\log d\right). \tag{12.8}$$