CONVERGENCE OF ZO-GD WITH SPHERE SMOOTHING

Kun Yuan

October 18, 2023

1 Algorithm

Consider the following optimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x) \tag{1}$$

in which we cannot access the gradient information of $f(x)$. The zeroth-order gradient descent algorithm iterates as follows:

$$g_k = \frac{d}{\tau}(f(x_k + \tau u_k) - f(x_k))u_k, \quad u_k \sim \mathcal{U}(\mathbb{S}^{d-1}(0,1)), \tag{2a}$$

$$x_{k+1} = x_k - \gamma g_k. \tag{2b}$$

Since u_k is a random variable for any $k = 0, 1, \cdots$, variables g_k and x_k are also random during the entire iteration process.

2 Sphere smoothing properties

Lemma 2.1. If $f(x)$ is L-smooth, it holds for any $x_k \in \mathbb{R}^d$ that

$$\|\nabla f(x_k) - \mathbb{E}_u[g_k]\| \leq L\tau \tag{3}$$

Lemma 2.2. If $f(x)$ is L-smooth, it holds for any $x_k \in \mathbb{R}^d$ that

$$\mathbb{E}_u\|g(x_k)\|^2 \leq 2d\|\nabla f(x_k)\|^2 + \frac{\tau^2 L^2 d^2}{2} \tag{4}$$

3 Convergence analysis

3.1 Non-convex analysis

Theorem 3.1. Assume $f(x)$ is L-smooth and $d \geq 8$. If $\gamma = 1/(4dL)$, ZO-GD with sphere smoothing converges as follows:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(x_k)\|^2 \leq \frac{16dL(f(x_0) - f(x^\star))}{K+1} + \frac{dL^2\tau^2}{2}. \tag{5}$$

Proof. Since $f(x)$ is L-smooth, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma\langle \nabla f(x_k), g_k\rangle + \frac{L\gamma^2}{2}\|g_k\|^2 \end{aligned} \tag{6}$$

We introduce the filtration $\mathcal{F}_k = \{x_k, u_{k-1}, x_{k-1}, \cdots, u_0, x_0\}$ to facilitate the analysis. By taking conditional expectations over both sides of inequality (6), we have

$$\begin{aligned} \mathbb{E}_u[f(x_{k+1})|\mathcal{F}_k] &\leq f(x_k) - \gamma\langle \nabla f(x_k), \mathbb{E}_u[g_k]\rangle + \frac{L\gamma^2}{2}\mathbb{E}_u\|g_k\|^2 \\ &= f(x_k) - \frac{\gamma}{2}\|\nabla f(x_k)\|^2 - \frac{\gamma}{2}\|\mathbb{E}_u[g_k]\|^2 + \frac{\gamma}{2}\|\mathbb{E}_u[g_k] - \nabla f(x_k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_u\|g_k\|^2 \\ &\leq f(x_k) - \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{\gamma L^2\tau^2}{2} + dL\gamma^2\|\nabla f(x_k)\|^2 + \frac{\tau^2 L^3 d^2\gamma^2}{4} \\ &= f(x_k) - \gamma(\frac{1}{2} - dL\gamma)\|\nabla f(x_k)\|^2 + \frac{\gamma L^2\tau^2}{2}(1 + \frac{\gamma d^2 L}{2}). \end{aligned} \tag{7}$$

If $\gamma \leq 1/(4dL)$, the above inequality becomes

$$\mathbb{E}_u[f(x_{k+1})|\mathcal{F}_k] \leq f(x_k) - \frac{\gamma}{4}\|\nabla f(x_k)\|^2 + \frac{\gamma L^2\tau^2}{2}(1 + \frac{d}{8}). \tag{8}$$

Taking expectations over the filtration $\mathcal{F}_k$, we have

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \frac{\gamma}{4}\mathbb{E}\|\nabla f(x_k)\|^2 + \frac{\gamma L^2\tau^2}{2}(1 + \frac{d}{8}), \tag{9}$$

which implies that

$$\begin{aligned}\mathbb{E}\|\nabla f(x_k)\|^2 &\le \frac{4}{\gamma}\left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]\right) + 2L^2\tau^2(1+\frac{d}{8}) \\ &\le \frac{4}{\gamma}\left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]\right) + \frac{dL^2\tau^2}{2} \end{aligned} \tag{10}$$

where the last inequality holds when $d \ge 8$. Taking the average over $k = 0, \cdots, K$, we have

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(x_k)\|^2 \le \frac{4(f(x_0) - f(x^\star))}{\gamma(K+1)} + \frac{dL^2\tau^2}{2}. \tag{11}$$

Setting $\gamma = 1/(4dL)$, we achieve the final result. □

Remark. When we use time-varying τ_k in Algorithm 2 such that $\sum_{k=0}^{K}\tau_k^2 \le R^2$, ZO-GD will converge to the exact stationary point at rate

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}\|\nabla f(x_k)\|^2 \le \frac{16dL(f(x_0) - f(x^\star))}{K+1} + \frac{dL^2R^2}{2(K+1)}. \tag{12}$$

We leave the proof as the exercise.

3.2 Strongly-convex analysis

Theorem 3.2. Assume $f(x)$ is L-smooth, μ-strongly convex, and $d \ge 8$. If $\gamma = 1/(4dL)$, ZO-GD with sphere smoothing converges as follows:

$$\mathbb{E}[f(x_k)] - f(x^\star) \le (1 - \frac{\mu}{8dL})^k\big(\mathbb{E}[f(x_0)] - f(x^\star)\big) + \frac{dL^2\tau^2}{4\mu}. \tag{13}$$

Proof. Since $f(x)$ is L-smooth and μ-strongly convex, we have

$$\|\nabla f(x)\|^2 \ge 2\mu(f(x) - f(x^\star)). \tag{14}$$

Its proof can be referred to Eq.(14) in our notes for Chapter 5. Substituting the above inequality into (15), we have

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - \frac{\gamma\mu}{2}\mathbb{E}[f(x_k) - f(x^\star)] + \frac{\gamma L^2\tau^2}{2}(1+\frac{d}{8}). \tag{15}$$

Subtracting $f(x^\star)$ from both sides of the above inequality, we have

$$\mathbb{E}[f(x_{k+1})] - f(x^\star) \le (1 - \frac{\gamma\mu}{2})\big(\mathbb{E}[f(x_k)] - f(x^\star)\big) + \frac{\gamma dL^2\tau^2}{8} \tag{16}$$

where we also used the assumption that $d \ge 8$. Keep iterating the above inequality, we have

$$\mathbb{E}[f(x_k)] - f(x^\star) \le (1 - \frac{\gamma\mu}{2})^k\big(\mathbb{E}[f(x_0)] - f(x^\star)\big) + \frac{dL^2\tau^2}{4\mu}. \tag{17}$$

Setting $\gamma = 1/(4dL)$ achieves the final result. □