Zero Order Method SGD Convergence

Dhruv Malik

March 22nd 2018

Setup. Consider a function $f(x) : \mathbb{R}^d \mapsto \mathbb{R}$ which has L-Lipschitz gradients:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2 \tag{1}$$

and also satisfies the PL Inequality:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu(f(x) - f(x^*)) \tag{2}$$

Assume we are optimizing over some bounded region and define $B = \sup_{x,y} \|x - y\|_2$ for all x, y in the region of optimization.

Method. Consider the following zero order optimization method: we randomly sample a direction r from a uniform distribution over the hypersphere with unit radius which is centered at the origin. Define w to be the following random variable:

$$w = \begin{cases} r, & \text{if } r^\top \nabla f(x) \geq 0 \\ -r, & \text{if } r^\top \nabla f(x) < 0 \end{cases} \tag{3}$$

Then, follow the update rule $x_{t+1} = x_t - \eta_t w$. We will show convergence in expectation.

Lemma 1. *If the unit vector w is selected using the procedure above, then:*

$$\mathbb{E}[w^\top \nabla f(x)] = O\left(\frac{\|\nabla f(x)\|_2}{\sqrt{d}}\right) \tag{4}$$

Proof. We first show an upper bound. Define the unit vector $v = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^d$. There exists an orthogonal transformation U such that $\|\nabla f(x)\|_2 Uv = \nabla f(x)$. Also, by the rotational invariance of the uniform distribution, we have that the previous definition for w is equivalent to:

$$w = \begin{cases} Ur, & \text{if } r^\top v \geq 0 \\ -Ur, & \text{if } r^\top v < 0 \end{cases} \tag{5}$$

Let r_1 the first entry of the vector r. Now:

$$\begin{aligned} \mathbb{E}[w^\top \nabla f(x)] &= \|\nabla f(x)\|_2 \, \mathbb{E}[w^\top (Uv)] && (6) \\ &= \|\nabla f(x)\|_2 \, \mathbb{E}[w^\top (Uv)] && (7) \\ &= \|\nabla f(x)\|_2 \, \mathbb{E}[\operatorname{sgn}(r_1) \cdot r_1] && (8) \\ &= \|\nabla f(x)\|_2 \, \mathbb{E}[|r_1|] && (9) \end{aligned}$$

Further, we have that:

$$(\mathbb{E}[|r_1|])^2 \leq \mathbb{E}[(r_1)^2] \tag{10}$$
$$= \frac{1}{d}\mathbb{E}[\sum_{i=1}^{d}(r_i)^2] \tag{11}$$
$$= \frac{1}{d} \tag{12}$$

This then implies that $\mathbb{E}[w^\top \nabla f(x)] \leq \frac{\|\nabla f(x)\|_2}{\sqrt{d}}$.

We now show a lower bound. For $i = 1 \ldots d$, define the random variables $Z_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. Define Z to be the vector $(Z_1 \ldots Z_d)$. Note that the random variable $\frac{Z}{\|Z\|_2}$ defines a uniform distribution over $\mathcal{S}^{d-1}$, the unit sphere in d dimensions. Define $\mathcal{E}$ to be the event when $\|Z\|_2 \leq 2\sqrt{d}$, $\mathcal{E}^C$ is the complement of this event, and Y is the random variable which is defined on $\mathcal{E}$ and $\mathcal{E}^C$. Then we have:

$$\mathbb{E}[|r_1|] = \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\right] \tag{13}$$
$$= \mathbb{E}\left[\mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\right]\middle| Y\right] \tag{14}$$
$$= \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle|\mathcal{E}\right]P(\mathcal{E}) + \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle|\mathcal{E}^C\right]P(\mathcal{E}^C) \tag{15}$$
$$= \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle| \|Z\|_2 \leq 2\sqrt{d}\right]P(\|Z\|_2 \leq 2\sqrt{d}) + \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle| \|Z\|_2 > 2\sqrt{d}\right]P(\|Z\|_2 > 2\sqrt{d}) \tag{16}$$
$$\geq \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle| \|Z\|_2 \leq 2\sqrt{d}\right]P(\|Z\|_2 \leq 2\sqrt{d}) \tag{17}$$

Now, note that $P(\|Z\|_2 \leq 2\sqrt{d}) = P(\|Z\|_2^2 \leq 4d)$. $\|Z\|_2^2$ has the chi-squared distribution, which is known to be subexponential with parameters $(2\sqrt{d}, 4)$. So, $P(\|Z\|_2 \leq 2\sqrt{d}) \geq \frac{1}{2}$. Therefore, we have that:

$$\mathbb{E}[|r_1|] \geq \mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle| \|Z\|_2 \leq 2\sqrt{d}\right]P(\|Z\|_2 \leq 2\sqrt{d}) \tag{18}$$
$$\geq \frac{1}{2}\mathbb{E}\left[\frac{|Z_1|}{\|Z\|_2}\middle| \|Z\|_2 \leq 2\sqrt{d}\right] \tag{19}$$
$$\geq \frac{1}{4\sqrt{d}}\mathbb{E}[|Z_1|] \tag{20}$$
$$= O(\frac{1}{\sqrt{d}}) \tag{21}$$

We therefore have that $\mathbb{E}[w^\top \nabla f(x)] = \|\nabla f(x)\|_2 \, \mathbb{E}[|r_1|] \geq O(\frac{\|\nabla f(x)\|_2}{\sqrt{d}})$. Combining the lower and upper bounds gives us our result. □

Lemma 2. *At any point x we have that $-\|\nabla f(x)\|_2 \geq -LB$.*

Proof. The conditions imply that $\nabla f(x) = 0$ if and only if x is a global minimizer of the function f. Let x^* be the global minimizer of f. Then for any x:

$$\|\nabla f(x)\|_2 = \|\nabla f(x) - \nabla f(x^*)\|_2 \tag{22}$$
$$\leq L \|x - x^*\|_2 \tag{23}$$
$$\leq LB \tag{24}$$

Then we have that $-\|\nabla f(x)\|_2 \geq -LB$. $\square$

Theorem 1. *Define the variable step size $\eta_t = \alpha_t BL$, where $\alpha_t = \frac{(2k+1)\sqrt{d}}{2\mu(k+1)^2}$. If we use the optimization method given above for $f(x) : \mathbb{R}^d \mapsto \mathbb{R}$, then we have an $O(\frac{d}{\epsilon})$ convergence rate as follows:*

$$\mathbb{E}\Big[f(x_t) - f(x^*)\Big] \leq \frac{L^3 B^2 d}{2\mu^2 t} \tag{25}$$

Proof. For simplicity, I assume that the result given in Lemma 1 exactly holds, i.e. that $\mathbb{E}[w^\top \nabla f(x)] = \frac{\|\nabla f(x)\|_2}{\sqrt{d}}$. Given an x_t at any timestep t:

$$\mathbb{E}\Big[f(x_{t+1}) - f(x_t)\Big] \leq \mathbb{E}\Big[\nabla f(x)^\top (x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|_2^2\Big] \tag{26}$$
$$= \mathbb{E}\Big[\nabla f(x)^\top (-\eta_t w) + \frac{L}{2}\|-\eta_t w\|_2^2\Big] \tag{27}$$
$$= -\eta_t \frac{\|\nabla f(x)\|_2}{\sqrt{d}} + \frac{L}{2}\eta_t^2 \tag{28}$$
$$= -\alpha_t BL \frac{\|\nabla f(x)\|_2}{\sqrt{d}} + \frac{L}{2}(\alpha_t BL)^2 \tag{29}$$
$$\leq -\alpha_t \frac{\|\nabla f(x)\|_2^2}{\sqrt{d}} + \frac{L^3 B^2}{2}\alpha_t^2 \tag{30}$$
$$\leq -\alpha_t \frac{2\mu}{\sqrt{d}}(f(x_t) - f(x^*)) + \frac{L^3 B^2}{2}\alpha_t^2 \tag{31}$$

This then implies that:

$$\mathbb{E}\Big[f(x_{t+1}) - f(x^*)\Big] \leq (1 - \alpha_t \frac{2\mu}{\sqrt{d}})(f(x_t) - f(x^*)) + \frac{L^3 B^2}{2}\alpha_t^2 \tag{32}$$
$$= \frac{t^2}{(t+1)^2}(f(x_t) - f(x^*)) + \frac{L^3 B^2 d(2t+1)^2}{8\mu^2 (t+1)^4} \tag{33}$$

If we define $C^2 = L^2 B^2 d$, then the remainder of the analysis is identical to that shown in Schmidt et al., starting from the bottom of page 6. $\square$