CHAPTER 5. ZEROTH-ORDER OPTIMIZATION

Yilong Song

October 17, 2023

1 Application Scenarios

1.1 Adversarial example generation for deep neural network

Let $\boldsymbol{\theta}$ be a given DNN, from which we have complete access to the corresponding output of our input. However, the internal states/configurations and the operating mechanism of DL systems are not revealed to us. (Which calls a black-box)
The adversarial example generation problem can be described as:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(\boldsymbol{z}+\boldsymbol{x};\boldsymbol{\theta})+\lambda g(\boldsymbol{x}) \\ s.t. \quad & \|\boldsymbol{x}\|\leq\epsilon, \end{aligned}$$

where $\boldsymbol{z}$ denotes the former legitimate example and $\boldsymbol{z}+\boldsymbol{x}$ denotes an adversarial example, with the perturbation $\boldsymbol{x}$, f denotes the attack loss function, and $\lambda g(\boldsymbol{x})$ is a regularization function. Subject to the black-box $\boldsymbol{\theta}$, the accurate expression of f is not accessible, where zeroth-order methods apply well. [1]

1.2 Online sensor management

The problem could be described as:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^d} \quad & \frac{1}{T}\sum_{t=1}^{T}\left[-\log\det\left(\sum_{i=1}^{d}x_i\boldsymbol{a}_{i,t}\boldsymbol{a}_{i,t}^{\top}\right)\right] \\ \text{s.t.} \quad & \mathbf{1}^{\top}\boldsymbol{x}=m_0, \quad \mathbf{0}\leq\boldsymbol{x}\leq\mathbf{1}, \end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the optimization variable, d is the number of sensors, $\boldsymbol{a}_{i,t}$ is the observation coefficient of sensor i at time t, and m_0 is the number of selected sensors.
To calculate the gradient, we need the inverses of large matrices, which is costly.

2 Problem formulation

We list several situations where the first-order methods lose efficacy:

- Gradients are costly to compute.
- A black-box which keeps the accurate function expression invisible.
- The observed function value contains noise, which makes gradient not representative.

3 Standard Finite Differences

3.1 Gradient Estimator

Review: The partial derivative of f can be expressed as:

$$\frac{\partial f(\boldsymbol{x}_0)}{\partial x_i} = \lim_{t \to 0} \frac{f(\boldsymbol{x}_0 + t\boldsymbol{e}_i) - f(\boldsymbol{x}_0)}{t}, \quad i = 1, 2, \ldots, d. \tag{1}$$

and the gradient is:

$$\nabla f(\boldsymbol{x}_0) = \left(\frac{\partial f(\boldsymbol{x}_0)}{\partial x_1}, \ldots, \frac{\partial f(\boldsymbol{x}_0)}{\partial x_d} \right)^\top = \sum_{i=1}^{d} \frac{\partial f(\boldsymbol{x}_0)}{\partial x_i} \boldsymbol{e}_i \tag{2}$$

However, the operation of limit is only theoretically feasible, for practice, an excessively small t for numerical computation brings a large scale of error. It is workable that we use a finite t instead of the limit of t to estimate gradient.
So we have a preliminary gradient estimator:

$$\hat{\nabla} f(x_0)_t = \sum_{i=1}^{d} \frac{f(\boldsymbol{x}_0 + t\boldsymbol{e}_i) - f(\boldsymbol{x}_0)}{t} \boldsymbol{e}_i \tag{3}$$

In order to obtain an estimator of the gradient at one point, we need $d + 1$ samples of function value, namely $\mathcal{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_0 + t\boldsymbol{e}_1, \ldots \boldsymbol{x}_0 + t\boldsymbol{e}_d\}$. The estimator above is called **forward difference**. Sometimes we also use a "two-side" estimator, which we call **central difference**:

$$\hat{\nabla} f(x_0)_t = \sum_{i=1}^{d} \frac{f(\boldsymbol{x}_0 + t\boldsymbol{e}_i) - f(\boldsymbol{x}_0 - t\boldsymbol{e}_i)}{2t} \boldsymbol{e}_i \tag{4}$$

This method equals to a mean value on two opposite directions around $\boldsymbol{x}_0$, namely $\mathcal{X} = \{\boldsymbol{x}_0 \pm t\boldsymbol{e}_1, \ldots \boldsymbol{x}_0 \pm t\boldsymbol{e}_d\}$, it requires $2d$ samples for one point, but usually this method is more accurate.

3.2 Error Analysis

Since a gradient without any limits could be inscrutable, we need some assumption for the gradient of f if we want to bound the error, although the gradient is not achievable in practice.

Assumption 3.1. (Lipschitz continuity of the gradients of f)
The function f is continuously differentiable, and the gradient of f is L-Lipschitz continuous for all $x \in \mathbb{R}^d$. Namely,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d \tag{5}$$

Notation. This assumption ensures L-smooth of f, and it also guarantees that f is not too concave.

Theorem 3.2. Let $g(x)$ denote the forward finite difference approximation to the gradient $\nabla f(x)$, if $\nabla f(x)$ is L-Lipschitz continuous, then for all $x \in \mathbb{R}^d$,

$$\|g(x) - \nabla f(x)\| \leq \frac{\sqrt{d}Lt}{2} \tag{6}$$

Proof.

$$tg(x)^\top \boldsymbol{e}_i = f(x + t\boldsymbol{e}_i) - f(x)$$
$$= \int_0^t \frac{\partial}{\partial x_i} f(x + u\boldsymbol{e}_i)\mathrm{d}u = \int_0^t \boldsymbol{e}_i^\top \nabla f(x + u\boldsymbol{e}_i)\mathrm{d}u$$

So,

$$\begin{aligned}
&t|\boldsymbol{e}_i^\top (g(x) - \nabla f(x))| \\
&= |\int_0^t \boldsymbol{e}_i^\top \nabla f(x + u\boldsymbol{e}_i)\mathrm{d}u - t\boldsymbol{e}_i^\top \nabla f(x)| \\
&= |\int_0^t \boldsymbol{e}_i^\top (\nabla f(x + u\boldsymbol{e}_i) - \nabla f(x))\mathrm{d}u| \\
&\leq \int_0^t \|\nabla f(x + u\boldsymbol{e}_i) - \nabla f(x)\|\mathrm{d}u \\
&\leq \int_0^t Lu\mathrm{d}u = \frac{Lt^2}{2}
\end{aligned}$$

So for every entry of $g(x) - \nabla f(x)$,

$$|(g(x) - \nabla f(x))_i| \leq \frac{Lt}{2}$$

then for the d-dimension vector,

$$\|g(x) - \nabla f(x)\| \leq \frac{\sqrt{d}Lt}{2}$$

□

Observation. It's obvious that the error would be smaller if we choose a smaller t, which will, however, bring in higher computational error.

3.3 Zeroth-order gradient descent with forward difference

With the finite difference estimators introduced above, we have a zeroth-order gradient descent (ZO-GD) method as follows

$$\boldsymbol{g}(\boldsymbol{x}_k) = \sum_{i=1}^{d} \frac{f(\boldsymbol{x}_k + t\boldsymbol{e}_i) - f(\boldsymbol{x}_k)}{t} \boldsymbol{e}_i \tag{7}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma \boldsymbol{g}(\boldsymbol{x}_k) \tag{8}$$

3.3.1 Convergence in the non-convex scenario

Theorem 3.3. Assume $f(\boldsymbol{x})$ to be L-smooth. If we set $\gamma = 1/L$, ZO-GD converges as follows

$$\frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2L(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star))}{K+1} + \frac{dL^2t^2}{4}. \tag{9}$$

Proof. Since $f(\boldsymbol{x})$ is L-smooth, we have

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 \\ &= f(\boldsymbol{x}_k) - \gamma \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{g}(\boldsymbol{x}_k) \rangle + \frac{\gamma^2 L}{2} \|\boldsymbol{g}(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - \frac{\gamma}{2} \|\nabla f(\boldsymbol{x}_k)\|^2 - \frac{\gamma}{2}(1 - L\gamma) \|\boldsymbol{g}(\boldsymbol{x}_k)\|^2 + \frac{\gamma}{2} \|\boldsymbol{g}(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)\|^2 \\ &\leq f(\boldsymbol{x}_k) - \frac{\gamma}{2} \|\nabla f(\boldsymbol{x}_k)\|^2 + \frac{\gamma d L^2 t^2}{8} \end{aligned} \tag{10}$$

where the last inequality holds due to Theorem 3.2 and $\gamma \leq 1/L$. Taking the average of the above inequality over $k = 0, 1, \cdots, K$, we achieve

$$\frac{1}{K+1} \sum_{k=0}^{K} \|\nabla f(\boldsymbol{x}_k)\|^2 \leq \frac{2(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star))}{\gamma(K+1)} + \frac{dL^2t^2}{4}. \tag{11}$$

If we let $\gamma = 1/L$, then we prove the result. □

3.3.2 Convergence in the strongly-convex scenario

Theorem 3.4. Assume $f(\boldsymbol{x})$ to be L-smooth and μ-strongly convex. If we set $\gamma = 1/L$, ZO-GD converges as follows

$$f(\boldsymbol{x}_K) - f(\boldsymbol{x}^\star) \leq (1 - \frac{\mu}{L})^K (f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)) + \frac{dL^2t^2}{8\mu} \tag{12}$$

Proof. Since $f(x)$ is μ-strongly convex, it holds from Theorem 3.8 of our notes "Ch0" that

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2\mu} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2. \tag{13}$$

Let $\boldsymbol{y} = \boldsymbol{x}_k$ and $\boldsymbol{x} = \boldsymbol{x}^\star$, we have

$$\|\nabla f(\boldsymbol{x}_k)\|^2 \geq 2\mu \big(f(\boldsymbol{x}^k) - f(\boldsymbol{x}^\star)\big). \tag{14}$$

Substituting the above inequality into (10), we achieve

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq (1 - \mu\gamma)(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)) + \frac{\gamma d L^2 t^2}{8} \tag{15}$$

Keep iterating the above inequality, we have

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq (1 - \mu\gamma)^{k+1}(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)) + \frac{d L^2 t^2}{8\mu} \tag{16}$$

Setting $\gamma = 1/L$ acheives the final result. □

4 Linear Interpolation

4.1 Gradient Estimator

Standard Finite Differences use the sample set $\mathcal{X} = \{x, x + t\boldsymbol{e}_1, \ldots, x + t\boldsymbol{e}_d\}$. Although gradients are always expressed in terms of entries when we need to calculate them, it's not necessary that we should always estimate the gradient along the directions of d axes. In fact, we can use a group of d linearly independent vectors $\{\boldsymbol{u}_i\}_{i=1}^d$ in stead of standard unit vectors. So the sample set now becomes $\mathcal{X} = \{x, x + t\boldsymbol{u}_1, \ldots \boldsymbol{u}_d\}$. It's notable that u_i needn't be a unit vector, but since the scale of u_i is arbitrary, how can we construct a proper gradient estimator? We need the tool of **linear interpolation**.[2]
As Taylor's formula shows,

$$f(x + \Delta x) = f(x) + \nabla f(x)^\top \cdot \Delta x + o(\|\Delta x\|) \tag{17}$$

We extract the linear part,

$$m(x + \Delta x) = f(x) + \nabla f(x)^\top \cdot \Delta x \tag{18}$$

It's a d-dimension hyperplane in $(d+1)$-dimension space, which we need $d+1$ points to determine when $\nabla f(x)$ is unknown. Precisely, we have $d+1$ points in the sample set above. Namely, let $m(x + t\boldsymbol{u}_i) = f(x + t\boldsymbol{u}_i)$. We still use $\boldsymbol{g}(x)$ to represent the undetermined coefficients, the equation could be expressed as

$$\begin{pmatrix} f(x + t\boldsymbol{u}_1) - f(x) \\ \vdots \\ f(x + t\boldsymbol{u}_d) - f(x) \end{pmatrix} = t \begin{pmatrix} \boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_d^\top \end{pmatrix} \boldsymbol{g}(x) \tag{19}$$

or we simplify it as

$$F_{\mathcal{X}} = tQ_{\mathcal{X}}\boldsymbol{g}(x) \tag{20}$$

So

$$\begin{aligned}\boldsymbol{g}(x) &= \frac{1}{t}Q_{\mathcal{X}}^{-1}F_{\mathcal{X}} \\ &\triangleq \frac{1}{t}(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d)\begin{pmatrix} f(x+t\boldsymbol{u}_1)-f(x) \\ \vdots \\ f(x+t\boldsymbol{u}_d)-f(x)\end{pmatrix} \\ &= \sum_{i=1}^{d}\frac{f(x+t\boldsymbol{u}_i)-f(x)}{t}\boldsymbol{v}_i\end{aligned}$$

where $\{\boldsymbol{v}_i\}$ are a group of generated vectors of $\{\boldsymbol{u}_i\}$ and satisfy

$$\begin{pmatrix}\boldsymbol{u}_1^\top \\ \vdots \\ \boldsymbol{u}_n^\top\end{pmatrix}(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d) = \mathbf{I}_d \tag{21}$$

4.2 Error Analysis

Notation. For some $\boldsymbol{u}_i$ whose scale is excessively large, we tend to involve the scalar in t and proportionally shrink every $\boldsymbol{u}_i$. So without loss of generality, we define $\|\boldsymbol{u}_i\| \leq 1$ for all $i = 1, 2, \ldots$.

Theorem 4.1. Let $g(x)$ denote the forward linear interpolation approximation to the gradient $\nabla f(x)$, where $\mathcal{X} = \{x+t\boldsymbol{u}_1, \ldots x+t\boldsymbol{u}_d\}$ is a set of interpolation points such that $\max_{1\leq i\leq d}\|\boldsymbol{u}_i\| \leq 1$ and $Q_{\mathcal{X}}$ is non-singular. If $\nabla f(x)$ is L-Lipschitz continuous, then for all $x \in \mathbb{R}^d$,

$$\|g(x) - \nabla f(x)\| \leq \frac{\|Q_{\mathcal{X}}^{-1}\|_2\sqrt{d}Lt}{2} \tag{22}$$

Proof.

$$\begin{aligned}t\boldsymbol{g}(x)^\top\boldsymbol{u}_i &= f(x+t\boldsymbol{u}_i) - f(x) \\ &= \int_0^t \boldsymbol{u}_i^\top\nabla f(x+v\boldsymbol{u}_i)\mathrm{d}v\end{aligned}$$

So,

$$
\begin{aligned}
&t|\boldsymbol{u}_i^\top(g(x)-\nabla f(x))| \\
&=|\int_0^t \boldsymbol{u}_i^\top \nabla f(x+v\boldsymbol{u}_i)\mathrm{d}v - t\boldsymbol{u}_i^\top \nabla f(x)| \\
&=|\int_0^t \boldsymbol{u}_i^\top (\nabla f(x+v\boldsymbol{u}_i)-\nabla f(x))\mathrm{d}v| \\
&\leq \int_0^t \|\boldsymbol{u}_i\|\|\nabla f(x+v\boldsymbol{u}_i)-\nabla f(x)\|\mathrm{d}v \\
&\leq \int_0^t L\|\boldsymbol{u_i}\|^2 v\mathrm{d}v = \frac{\|\boldsymbol{u}_i\|^2 Lt^2}{2} \leq \frac{Lt^2}{2}
\end{aligned}
$$

which implies,

$$\|Q_{\mathcal{X}}(g(x)-\nabla f(x))\| \leq \frac{\sqrt{d}Lt}{2}$$

so,

$$\|g(x)-\nabla f(x)\| \leq \frac{\|Q_{\mathcal{X}}^{-1}\|_2\sqrt{d}Lt}{2}$$

□

The result shows that large $\|Q_{\mathcal{X}}^{-1}\|$ can cause large deviation of $\boldsymbol{g}(x)$ from $\nabla f(x)$. Thus it is desirable to select $\mathcal{X}$ in such a way that the condition number of $Q_{\mathcal{X}}^{-1}$ is small. It's the best when $Q_{\mathcal{X}}$ is orthonormal.

4.3 Central Difference

To make an analogue of central difference in standard finite difference, we use the sample set $\mathcal{X} = \{\boldsymbol{x}_0 \pm t\boldsymbol{u}_1, \dots \boldsymbol{x}_0 \pm t\boldsymbol{u}_d\}$, and similarly let

$$\boldsymbol{g}(x) = \sum_{i=1}^{d} \frac{f(x+t\boldsymbol{u}_i)-f(x-t\boldsymbol{u}_i)}{2t}\boldsymbol{v}_i \tag{23}$$

Here we use $2d$ points to determine a hyperplane. It's actually the average of two hyperplanes.

5 Gaussian Smoothness

5.1 Random Estimator

Up to now, we always need $d+1$ samples to estimate the gradient at one point, but if we choose just one direction for each iteration instead of d directions? We can certainly iterate through d selected directions during d iterations, but there is a more efficient method: random method.

To make the estimator unbiased, we need a central symmetrical variable $\boldsymbol{u}$ (e.g. Gauss distribution or uniform distribution on a sphere), and along this direction, we preliminarily

define the gradient estimator as

$$\boldsymbol{g}(x) = \frac{f(x + t\boldsymbol{u}) - f(x)}{t}\boldsymbol{u} \tag{24}$$

It's notable that the direction of this estimator depends completely on the arbitrary $\boldsymbol{u}$, so it is far from a good estimator for the true gradient. However, when $\boldsymbol{u}$ points at a bad direction, the factor before it may be small or negative. So it could hopefully be a proper estimator in terms of expectation.

5.2 Smooth Function

Sometimes the reason why we can't get the gradient is the noise, that is to say, the properties of f could be somewhat bad. But with the help of random method, we could construct a smooth version of f as

$$\hat{f}(x)_t \triangleq \mathbf{E}_{\boldsymbol{v}} f(x + t\boldsymbol{v}) \tag{25}$$

where t is called **smooth parameter**, and $\boldsymbol{v}$ is another central symmetrical variable. This smooth function could be understood as the mean of the function value around $f(x)$. Usually the smooth version of f owes better properties than f.

5.3 Bridge Lemma for Gaussian Smoothness

Now we specify $\boldsymbol{u} \sim N(0, \mathbf{I}), \quad \boldsymbol{v} \sim N(0, \mathbf{I})$, the following lemma bridges the gap between random estimator and smooth function, which ensures the feasibility of Gaussian smoothness.

Lemma 5.1. (bridge lemma I) If $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{u} \sim N(0, \mathbf{I})$, for $f \in \mathbb{C}^1$, a certain x, and a certain t,

$$\mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x} + t\boldsymbol{u}) - f(\boldsymbol{x})}{t} \cdot \boldsymbol{u} = \nabla \mathbf{E}_{\boldsymbol{u}} f(\boldsymbol{x} + t\boldsymbol{u}) \tag{26}$$

Proof. Because

$$\mathbf{E}_{\boldsymbol{u}} \tfrac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t} \cdot \boldsymbol{u} = \begin{pmatrix} \mathbf{E}_{\boldsymbol{u}} \tfrac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t} \cdot u_1 \\ \vdots \\ \mathbf{E}_{\boldsymbol{u}} \tfrac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t} \cdot u_d \end{pmatrix},$$

for every entry, we have

$$
\begin{aligned}
&\mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t}\cdot u_1 = \mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t}\cdot u_1 \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d} \frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t}\cdot\frac{u_1}{(\sqrt{2\pi})^d}\mathrm{e}^{-\frac{u_1^2+\cdots+u_d^2}{2}}\mathrm{d}u_1\ldots\mathrm{d}u_d \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d-1} \frac{1}{(\sqrt{2\pi})^{d-1}}\mathrm{e}^{-\frac{u_2^2+\cdots+u_d^2}{2}}\mathrm{d}u_2\ldots\mathrm{d}u_d\int_{-\infty}^{+\infty}\frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t}\cdot\frac{u_1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u_1^2}{2}}\mathrm{d}u_1 \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d-1} \frac{1}{(\sqrt{2\pi})^{d-1}}\mathrm{e}^{-\frac{u_2^2+\cdots+u_d^2}{2}}\mathrm{d}u_2\ldots\mathrm{d}u_d\left(-\int_{-\infty}^{+\infty}\frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t}\cdot\frac{1}{\sqrt{2\pi}}\mathrm{d}\mathrm{e}^{-\frac{u_1^2}{2}}\right) \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d-1} \frac{1}{(\sqrt{2\pi})^{d-1}}\mathrm{e}^{-\frac{u_2^2+\cdots+u_d^2}{2}}\mathrm{d}u_2\ldots\mathrm{d}u_d\cdot \\
&\left(-\frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t}\cdot\frac{\mathrm{e}^{-\frac{u_1^2}{2}}}{\sqrt{2\pi}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u_1^2}{2}}\cdot\frac{\partial f(\boldsymbol{x}+t\boldsymbol{u})}{t\partial u_1}\cdot\frac{1}{\sqrt{2\pi}}\mathrm{d}u_1\right) \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d-1} \frac{1}{(\sqrt{2\pi})^{d-1}}\mathrm{e}^{-\frac{u_2^2+\cdots+u_d^2}{2}}\mathrm{d}u_2\ldots\mathrm{d}u_d\cdot\left(\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u_1^2}{2}}\cdot f_1'(\boldsymbol{x}+t\boldsymbol{u})\cdot\frac{1}{\sqrt{2\pi}}\mathrm{d}u_1\right) \\
=&\underbrace{\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}}_{d} f_1'(\boldsymbol{x}+t\boldsymbol{u})\cdot\frac{1}{(\sqrt{2\pi})^d}\mathrm{e}^{-\frac{u_1^2+\cdots+u_d^2}{2}}\mathrm{d}u_1\ldots\mathrm{d}u_d \\
=&\mathbf{E}_{\boldsymbol{u}} f_1'(\boldsymbol{x}+t\boldsymbol{u})
\end{aligned}
$$

Therefore,

$$
\begin{pmatrix}\mathbf{E}_{\boldsymbol{u}} f_1'(\boldsymbol{x}+t\boldsymbol{u})\\ \vdots \\ \mathbf{E}_{\boldsymbol{u}} f_d'(\boldsymbol{x}+t\boldsymbol{u})\end{pmatrix} = \mathbf{E}_{\boldsymbol{u}}\nabla f(\boldsymbol{x}+t\boldsymbol{u}) = \nabla\mathbf{E}_{\boldsymbol{u}} f(\boldsymbol{x}+t\boldsymbol{u}) \tag{27}
$$

□

Notation. $f\in\mathbb{C}^1$ is just a sufficient condition, the lemma holds among a wide range of functions, thanks to the good properties of normal distribution.

The lemma demonstrates that the expectation of our gradient estimator is precisely the gradient of the smooth function, which proves theoretically that our estimator is somehow an unbiased estimation of gradient. So the estimator could be denoted by

$$
\boldsymbol{g}(x) = \frac{f(x+t\boldsymbol{u})-f(x)}{t}\boldsymbol{u},\quad \boldsymbol{u}\sim N(0,\mathrm{I}) \tag{28}
$$

5.4 Two-point Estimator vs Single-point Estimator

From the proof above, we notice that $\mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t} \cdot \boldsymbol{u} = \mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x}+t\boldsymbol{u})}{t} \cdot \boldsymbol{u}$, since the expectation of the second term is zero. We name the former form **two-point estimator**, while the latter **single-point estimator**. These two forms of estimator are the same in terms of expectation, but different in practice. Obviously, single-point estimator will explode to ∞ as $t \to 0$, and it owes naturally a larger variance. The convergence analysis of single-point estimator is also different from the two-point estimator.

5.5 Error Analysis in Terms of Probability

The difference between our estimator and the true gradient can be divided into two parts:

$$\|\boldsymbol{g}(x) - \nabla f(x)\| \leq \|\nabla \hat{f}_t(x) - \nabla f(x)\| + \|\boldsymbol{g}(x) - \nabla \hat{f}_t(x)\| \tag{29}$$

The first term is the difference between the true function and the smooth function, which can be bounded as:

Lemma 5.2. if $\nabla f(x)$ is L-Lipschitz continuous, then for all $x \in \mathbb{R}^d$,

$$\|\nabla \hat{f}_t(x) - \nabla f(x)\| \leq \sqrt{d} L t \tag{30}$$

The second term is caused by the estimator. In fact, the covariance of $\boldsymbol{g}(x)$ could be enormously huge if we use only one sample. To reduce it, we can take N samples and calculate their mean.

$$\boldsymbol{g}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x + t\boldsymbol{u}_i) - f(x)}{t} \boldsymbol{u}_i, \quad \boldsymbol{u}_i \sim N(0, \mathrm{I}) \tag{31}$$

If we want to let the second term of error less than a determined r with probability $1 - \delta$, we need a large enough N. Considering both terms, the following theorem gives a lower bound of N.

Theorem 5.3. Suppose that $\nabla f(x)$ is L-Lipschitz continuous, if

$$N \geq \frac{3d}{\delta r^2} \left(3\|\nabla f(x)\|^2 + \frac{L^2 t^2}{4} (d+2)(d+4) \right) \tag{32}$$

then for all $x \in \mathbb{R}^d$ and $r>0$,

$$\|\boldsymbol{g}(x) - \nabla f(x)\| \leq \sqrt{d} L t + r \tag{33}$$

with probability at least $1 - \delta$.

5.6 Error Analysis in Terms of Expectation

Lemma 5.4. Suppose f is L-smooth, and let $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{I})$. Then,

$$\mathbf{E}_{\boldsymbol{u}}[\|\boldsymbol{g}(x)\|^2] \leq 2(d+2)\|\nabla f(x)\|^2 + \frac{t^2L^2}{2}d(d+2)(d+4) \tag{34}$$

Proof. We have

$$\begin{aligned}\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{g}(x)\|^2 &= \mathbf{E}_{\boldsymbol{u}}\left\|\frac{f(x+t\boldsymbol{u})-f(x)}{t}\cdot\boldsymbol{u}\right\|^2 \\ &= \frac{1}{t}\mathbf{E}_{\boldsymbol{u}}[\|f(x=t\boldsymbol{u})-f(x)\|^2\|\boldsymbol{u}\|^2],\end{aligned}$$

considering $(a+b)^2 \leq 2a^2+2b^2$, we have

$$\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{g}(x)\|^2 \leq \frac{2}{t^2}\mathbf{E}_{\boldsymbol{u}}[|f(x+t\boldsymbol{u})-f(x)-\langle\nabla f(x), t\boldsymbol{u}\rangle|^2\|\boldsymbol{u}\|^2] + 2\mathbf{E}_{\boldsymbol{u}}[|\nabla f(x),\boldsymbol{u}\rangle|^2\|\boldsymbol{u}\|^2].$$

First we consider the second term, note that

$$\mathbf{E}_{\boldsymbol{u}}[|\langle\nabla f(x),\boldsymbol{u}\rangle\|\boldsymbol{u}\|^2] = (\nabla f(x))^\top\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{u}\|^2\boldsymbol{u}\boldsymbol{u}^\top\nabla f(x). \tag{35}$$

When $i \neq j$, $\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{u}\|^2u_iu_j = 0$, for the independence and central symmetry of every variable. When $i=j$,

$$\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{u}\|u_i^2 = \mathbf{E}_{\boldsymbol{u}}(u_1^2+\cdots+u_d^2)u_i^2 = \mathbf{E}_{\boldsymbol{u}}u_1^4 + (d-1)(\mathbf{E}_{\boldsymbol{u}}u_1^2)^2 = 3+(d-1) = d+2. \tag{36}$$

Therefore,

$$\mathbf{E}_{\boldsymbol{u}}[\|\boldsymbol{u}\|^2\boldsymbol{u}\boldsymbol{u}^\top] = (d+2)\mathbf{I}. \tag{37}$$

$$2\mathbf{E}_{\boldsymbol{u}}[|\langle\nabla f(x),\boldsymbol{u}\rangle|^2\|x\|^2] = 2(d+2)\|\nabla f(x)\|^2 \tag{38}$$

Next we bound the first term: By Newton-Leibniz theorem,

$$f(x+t\boldsymbol{u})-f(x) = \int_0^t\frac{\mathrm{d}}{\mathrm{d}s}[f(x+s\boldsymbol{u})-f(x)]\mathrm{d}s = \int_0^t\langle\nabla f(x+s\boldsymbol{u}),\boldsymbol{u}\rangle\mathrm{d}s,$$

and thus,

$$\begin{aligned}|f(x+t\boldsymbol{u})-f(x)-\langle\nabla f(x),t\boldsymbol{u}\rangle| &= \int_0^t\langle\nabla f(x+s\boldsymbol{u})-\nabla f(x),\boldsymbol{u}\rangle\mathrm{d}s \\ &\leq \int_0^t\|\nabla f(x+s\boldsymbol{u})-\nabla f(x)\|\|\boldsymbol{u}\|\mathrm{d}s \\ &\leq \int_0^t Ls\|\boldsymbol{u}\|^2\mathrm{d}s = \frac{Lt^2}{2}\|\boldsymbol{u}\|^2\end{aligned}$$

We then get:

$$\frac{2}{t^2}\mathbf{E}_{\boldsymbol{u}}[|f(x+t\boldsymbol{u})-f(x)-\langle\nabla f(x),t\boldsymbol{u}\rangle|^2\|\boldsymbol{u}\|^2] \leq \frac{2}{t^2}\mathbf{E}_{\boldsymbol{u}}\left[\frac{L^2t^4}{4}\|\boldsymbol{u}^6\|\right] = \frac{L^2t^2}{2}d(d+2)(d+4),$$

where $\mathbf{E}_{\boldsymbol{u}}\|\boldsymbol{u}\|^6$ is the third moment of χ^2 distribution.

□

6 Sphere Smoothness

6.1 Bridge Lemma for Sphere Smoothness

Now we consider the uniform distribution on a d-dimension spherical surface, denoted by $\boldsymbol{u} \sim U(\mathbb{S}^{d-1}(0,1))$. Namely, $\boldsymbol{u}$ is a unit vector, and the direction is central symmetrical. When we use it to construct our gradient estimation, the according generated $\boldsymbol{v}$ is no longer the same as $\boldsymbol{u}$, but obeys uniform distribution in a unit ball, denoted by $\boldsymbol{v} \sim U(\mathbb{B}^d(0,1))$, coupled with a coefficient d. The following lemma proves it. (*We need the knowledge of integral on manifolds to accomplish our proof.) [3]

Lemma 6.1. (bridge lemma II) If $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{u} \sim U(\mathbb{S}^{d-1}(0,1))$, $\boldsymbol{v} \in \mathbb{R}^d$ and $\boldsymbol{v} \sim U(\mathbb{B}^d(0,1))$, for $f \in \mathbb{C}^1$, a certain x, and a certain t,

$$\mathbf{E}_{\boldsymbol{u}} \frac{f(\boldsymbol{x}+t\boldsymbol{u})-f(\boldsymbol{x})}{t} \cdot \boldsymbol{u} = \frac{1}{d}\nabla \mathbf{E}_{\boldsymbol{v}} f(\boldsymbol{x}+t\boldsymbol{v}) \tag{39}$$

Proof. *
If $d = 1$, then the fundamental theorem of calculus implies,

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_{-t}^{t} f(x+v)\mathrm{d}v = f(x+t) - f(x-t) \tag{40}$$

The d-dimensional generalization, which follows from Stoke's theorem, is,

$$\nabla \int_{t\mathbb{B}} f(x+\boldsymbol{v})\mathrm{d}\boldsymbol{v} = \int_{t\mathbb{S}} f(x+\boldsymbol{u})\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\mathrm{d}\boldsymbol{u} \tag{41}$$

By definition,

$$\mathbf{E}f(x+t\boldsymbol{v}) = \frac{\int_{t\mathbb{B}} f(x+\boldsymbol{v})\mathrm{d}\boldsymbol{v}}{\mathrm{vol}_d(t\mathbb{B})} \tag{42}$$

Similarly,

$$\mathbf{E}f(x+t\boldsymbol{u})\cdot\boldsymbol{u} = \frac{\int_{t\mathbb{S}} f(x+\boldsymbol{u})\cdot\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\mathrm{d}\boldsymbol{u}}{\mathrm{vol}_{d-1}(t\mathbb{S})} \tag{43}$$

Noticed that $\mathrm{vol}_d(t\mathbb{B})$ is $\frac{t}{d}$ times of $\mathrm{vol}_{d-1}(t\mathbb{S})$, then the lemma follows. □

So an unbiased estimator of sphere smoothness should be:

$$\boldsymbol{g}(x) = \frac{f(x+t\boldsymbol{u})-f(x)}{t}\cdot d\cdot\boldsymbol{u}, \quad \boldsymbol{u} \sim U(\mathbb{S}^{d-1}(0,1)) \tag{44}$$

where d is an extra coefficient.

6.2 Error Analysis in Terms of Probability

To make an analogue, we also use a minibatch of samples as:

$$\boldsymbol{g}(x) = \frac{d}{N}\sum_{i=1}^{N}\frac{f(x+t\boldsymbol{u})-f(x)}{t}\cdot\boldsymbol{u} \quad \boldsymbol{u}\sim U(\mathbb{S}^{d-1}(0,1)) \tag{45}$$

And the similar theorem is as follows

Theorem 6.2. Suppose that $\nabla f(x)$ is L-Lipschitz continuous, if

$$N \geq \left[\frac{6d^2}{r^2}\left(\frac{\|\nabla f(x)\|^2}{d}+\frac{L^2t^2}{4}\right)+\frac{2d}{3r}\left(2\|\nabla f(x)\|+Lt\right)\right]\log\frac{d+1}{\delta} \tag{46}$$

then for all $x\in\mathbb{R}^d$ and $r>0$,

$$\|\boldsymbol{g}(x)-\nabla f(x)\| \leq Lt+r \tag{47}$$

with probability at least $1-\delta$.

6.3 Error Analysis in Terms of Expectation

Lemma 6.3. Suppose f is L-smooth, and let $\boldsymbol{u}\sim U(\mathbb{S}^{d-1}(0,1))$. Then,

$$\mathbf{E}_{\boldsymbol{u}}[\|\boldsymbol{g}(x)\|^2] \leq 2d\|\nabla f(x)\|^2+\frac{t^2L^2d^2}{2} \tag{48}$$

The proof is similar to the parallel lemma for Gaussian smoothness, but we need the fact that:

$$\mathbf{E}_{\boldsymbol{u}}[\boldsymbol{u}\boldsymbol{u}^\top] = \frac{1}{d}\mathbf{I}, \quad \boldsymbol{u}\sim U(\mathbb{S}^{d-1}(0,1)), \tag{49}$$

which also demands the integral on a manifold.

6.4 Convergence Analysis in convex situation

Now we assume that $f(x)$ is convex and L-smooth, we analyze the convergence rate of Sphere Smoothness. With the assumption of convexity, the error between f and the smooth version of f has nothing to do with the number of dimension any more. [4]

Lemma 6.4. Suppose f is convex and L-smooth. Let $\boldsymbol{u}\sim N(\mathbf{0},\mathbf{I})$, and $\hat{f}_t$ denote the corresponding smooth version of f. Then $\hat{f}_t$ is convex, L-smooth and satisfies

$$f(x) \leq \hat{f}_t(x) \leq f(x)+\frac{Lt^2}{2} \tag{50}$$

and

$$\|\nabla\hat{f}_t(x)-\nabla f(x)\| \leq Lt \tag{51}$$

Proof. The convexity of $\hat{f}_t$ follows by noting that

$$\begin{aligned}\hat{f}_t(\theta x_1 + (1-\theta)x_2) &= \mathbf{E}_{\boldsymbol{v}}[f(\theta x_1 + (1-\theta)x_2 + t\boldsymbol{v})] \\ &= \mathbf{E}_{\boldsymbol{v}}[f(\theta(x_1 + t\boldsymbol{v}) + (1-\theta)(x_2 + t\boldsymbol{v}))] \\ &\leq \mathbf{E}_{\boldsymbol{v}}[\theta f(x_1 + t\boldsymbol{v} + (1-\theta)f(x_2 + t\boldsymbol{v})] \\ &= \theta\hat{f}_t(x_1) + (1-\theta)\hat{f}_t(x_2)\end{aligned}$$

for any $\theta \in [0,1]$ and any x_1, x_2. To show the L-smoothness of $\hat{f}_t$, let $x_1, x_2 \in \mathbb{R}^d$ be arbitrary, and we have

$$\begin{aligned}\|\nabla\hat{f}_t(x_1) - \nabla\hat{f}_t(x_2)\| &= \|\nabla\mathbf{E}_{\boldsymbol{v}}[f(x_1 + t\boldsymbol{v})] - \nabla\mathbf{E}_{\boldsymbol{v}}[f(x_2 + t\boldsymbol{v})]\| \\ &= \|\mathbf{E}_{\boldsymbol{v}}[\nabla f(x_1 + t\boldsymbol{v}) - \nabla f(x_2 + t\boldsymbol{v})]\| \\ &\leq \mathbf{E}_{\boldsymbol{v}}[\|\nabla f(x_1 + t\boldsymbol{v}) - \nabla f(x_2 + t\boldsymbol{v})\|] \\ &\leq \mathbf{E}_{\boldsymbol{v}}[L\|x_1 - x_2\|] = L\|x_1 - x_2\|.\end{aligned}$$

Now by the convexity and smoothness of f, we have

$$f(x) + \langle\nabla f(x), t\boldsymbol{v}\rangle \leq f(x + t\boldsymbol{v}) \leq f(x) + \langle\nabla f(x), t\boldsymbol{v}\rangle + \frac{L}{2}\|t\boldsymbol{v}\|^2 \tag{52}$$

Now we take the expectation with respect to $\boldsymbol{v} \sim N(\mathbf{0}, \mathbf{I})$, we have $\mathbf{E}_{\boldsymbol{v}}[\langle\nabla f(x), t\boldsymbol{v}\rangle] = 0$. Therefore,

$$f(x) \leq \mathbf{E}_{\boldsymbol{v}}[f(x + t\boldsymbol{v})] \leq f(x) + \frac{Lt^2}{2}\mathbf{E}_{\boldsymbol{v}}[\|\boldsymbol{v}\|^2], \tag{53}$$

which gives the first inequality. Now we regarding $\nabla\hat{f}_t$, we have

$$\begin{aligned}\|\nabla\hat{f}_t(x) - \nabla f(x)\| &= \|\nabla_x\mathbf{E}_{\boldsymbol{v}}[f(x + t\boldsymbol{v}) - f(x)]\| \\ &= \|\mathbf{E}_{\boldsymbol{v}}[\nabla_x f(x + t\boldsymbol{v}) - \nabla_x f(x)]\| \\ &\leq \mathbf{E}_{\boldsymbol{v}}[\|\nabla f(x + t\boldsymbol{v}) - \nabla f(x)\|] \\ &\leq \mathbf{E}_{\boldsymbol{v}}[L\|t\boldsymbol{v}\|] \leq Lt,\end{aligned}$$

□

We now make an iteration as follows, note that now t_k is not a constant, but a convergent series.

$$\boldsymbol{g}(x) = \frac{d}{t_k}(f(x + t_k\boldsymbol{u}) - f(x))\boldsymbol{u}, \quad \boldsymbol{u} \sim U(\mathbb{S}^{d-1}(0,1)) \tag{54}$$

$$x_{k+1} = x_k - \gamma\boldsymbol{g}(x) \tag{55}$$

We have the theorem for convergence rate for Sphere Smoothness as follows:

Theorem 6.5. Suppose f is convex and L-smooth, and has a minimizer $x \in \mathbb{R}^d$. Let $\gamma = c/(dL)$ for some $c \in (0,1)$, and let t_k be a positive sequence of smoothing parameter such that $\sum_{k=0}^{K} t_k^2 = R^2 \leq +\infty$. Then the zeroth-order optimization iteration achieves

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{E}[f(x_k) - f(x^*)] \leq \frac{dL\|x_0 - x^*\|^2}{c(1-c)(K+1)} + \frac{L}{2(1-c)}\left(1 + \frac{cd}{4}\right)\frac{\sum_{k=0}^{K} t_k^2}{K+1} \quad (56)$$

Proof.

$$\begin{aligned}\|x_{k+1} - x^*\|^2 =& \|x_k - x^* - \gamma \boldsymbol{g}(x)\|^2 \\ =& \|x_k - x^*\|^2 - 2\gamma\langle x_k - x^*, \boldsymbol{g}(x)\rangle + \gamma^2\|\boldsymbol{g}(x)\|^2.\end{aligned}$$

Using the lemma for $\|g(x)\|^2$ and the lemma for $\nabla \hat{f}_t$, we get

$$\mathbf{E}\langle x_k - x^*, \boldsymbol{g}(x)\rangle = \langle x_k - x^*, \nabla \hat{f}_{t_k}(x_k)\rangle, \quad (57)$$

$$\mathbf{E}\|\boldsymbol{g}(x)\|^2 \leq 2d\|\nabla f(x_k)\|^2 + \frac{t_k^2 L^2 d^2}{2}. \quad (58)$$

and consequently,

$$\mathbf{E}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma\langle x_k - x^*, \nabla \hat{f}_{t_k}(x_k)\rangle + 2\gamma^2 d\|\nabla f(x_k)\|^2 + \frac{\gamma^2 t_k^2 L^2 d^2}{2}. \quad (59)$$

Since $\hat{f}_{t_k}$ is convex, we see that

$$\hat{f}_{t_k}(x^*) - \hat{f}_{t_k}(x_k) \geq \langle \nabla \hat{f}_{t_k}(x_k), x^* - x_k\rangle, \quad (60)$$

and coupled with the lemma, we have

$$-\langle \nabla \hat{f}_{t_k}(x_k), x^* - x_k\rangle \leq \hat{f}_{t_k}(x^*) - \hat{f}_{t_k}(x_k) \leq f(x^*) - f(x_k) + \frac{Lt_k^2}{2}. \quad (61)$$

Moreover, since f is L-smooth and $\nabla f(x^*) = 0$, we have

$$\|\nabla f(x_k)\|^2 = \|\nabla f(x_k) - \nabla f(x^*)\|^2 \leq 2L(f(x_k) - f(x^*)). \quad (62)$$

Summarizing these results, we get

$$\mathbf{E}\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - 2\gamma(1 - 2\gamma dL)(f(x_k) - f(x^*)) + \gamma L t_k^2 + \frac{\gamma^2 t_k^2 L^2 d^2}{2}, \quad (63)$$

and by taking the total expectation, we can get

$$2\gamma(1 - 2\gamma dL)\mathbf{E}[f(x_k) - f(x^*)] \leq \mathbf{E}\|x_k - x^*\|^2 - \mathbf{E}\|x_{k+1} - x^*\|^2 + \gamma L t_k^2 + \frac{\gamma^2 t_k^2 L^2 d^2}{2}. \quad (64)$$

Now we take the telescoping sum and get

$$2\gamma(1-2\gamma dL)\sum_{k=0}^{K}\mathbf{E}[f(x_k)-f(x^*)] \leq \|x_0-x^*\|^2+\gamma L\left(1+\frac{\gamma L d^2}{2}\right)\sum_{k=0}^{K}t_k^2. \tag{65}$$

By taking $\gamma = c/(2dL)$ for some $c \in (0,1)$, we get the result:

$$\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{E}[f(x_k)-f(x^*)] \leq \frac{dL\|x_0-x^*\|^2}{c(1-c)(K+1)}+\frac{L}{2(1-c)}\left(1+\frac{cd}{4}\right)\frac{\sum_{k=0}^{K}t_k^2}{K+1} \tag{66}$$

$$= \frac{d}{K+1}\left(\frac{L\|x_0-x^*\|^2}{c(1-c)}+\frac{R^2L(c+4/d)}{8(1-c)}\right) \tag{67}$$

□

The result shows the convergence rate of zeroth-order optimization method with sphere smoothness is $O(d/K)$, which is still concerned with the number of dimension d.

References

[1] L. Sijia, C. Pin-Yu, K. Bhavya, Z. Gaoyuan, H. I. Alfred O., and V. Pramod K., "A primer on zeroth-orfer optimization in signal processing and machine learning - principals, recent advances, and applications," *Nonconvex Optimization for Signal Processing and Machine Learning*, pp. 43–54.

[2] B. Albert S., C. Liyuan, C. Krzysztof, and S. Katya, "A theoretical and empirical comparison of gradient approximations in derivative-free optimization," *Found Comput Math 22*, pp. 507–560.

[3] F. Abraham D., K. Adam Tauman, and M. H. Brendan, "Online convex optimization in the bandit setting: gradient descent without a gradient," *Proceedings of the sixteenth and annual ACM-SIAM symposium on discrete algorithms*, pp. 385–394.

[4] T. Yujie, "Introduction to zeroth-order optimization."