Zero-Order Optimization

Zening Liu

1. Introduction

Zero-order optimization methods, also called **derivative-free optimization** methods, deal with problems with no explicit access of gradients. These methods were among the first schemes introduced in the early days of the development of optimization theory [1] (see the references therein). They have an evident advantage over the first-order optimization methods: the computation of function values is always much simpler than the computation of gradients. However, since (1) zero-order optimization methods are much more difficult for theoretical investigation, (2) the possible convergence rate of these methods is far below the convergence rate of first-order optimization methods, (3) the technique called Fast Differentiation is developed for automatically computing the whole vector of the partial derivatives with at most four times larger computation complexity than the computation of the function value, if the function can be explicitly represented by a sequence of differentiable operations, the zero-order optimization methods were almost out of the sight, during the past several decades [1].

However, in the last years, we can see a restoration of the interest to these methods. These methods can find their place on the situations where explicit expressions of the gradients are difficult or infeasible to obtain. For example, only black-box procedures are available for computing the values of the functional characteristics of the problem, or resource limitations restrict the use of fast or automatic differentiation techniques, or privacy issues prohibit one from obtaining the gradients [1].

In zero-order optimization, the full gradient is typically approximated by using either a **one-point** or a **two-point** gradient estimator, which are usually called **zero-order oracles**. The former acquires a gradient estimate by querying the function at a single random location close to the point, and the latter computes a finite difference using two random function queries. Then, these gradient estimators are used to complete the update, instead of the real gradients. Recall that, for a one-dimension differentiable function $f : \mathbb{R} \to \mathbb{R}$, we can always use $\frac{f(x+\delta)-f(x-\delta)}{2\delta}$ to estimate the derivative of f at point x, i.e., $\frac{f(x+\delta)-f(x-\delta)}{2\delta} \approx f'(x)$.

Typically, the two-point gradient estimator has a lower variance and thus can improve the complexity bounds of zero-order algorithms [2], [3]. Thus, in this note, we mainly focus on the two-point oracles. In the rest of this note, we try to explain two things:

- What are zero-order oracles? We will summarize the up-to-date zero-order oracles introduced in different works in a unified way.
- Why are these zero-order oracles? We will explain the theoretical basis of these zero-order oracles, which are based on the various smoothing techniques, and summarize their properties under different levels of smoothness.

Notation: In this note, scalars are denoted by italic letters. Boldface lower-case letters denote vectors. Boldface upper-case letters denote matrixes. $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. $\mathbb{R}^d$ denotes the space of d-dimensional real-valued vectors. For a scalar-valued differential function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ denote its gradients and Hessians at point $\mathbf{x}$. Also, denote the ℓ_2-norm by $\|\cdot\|$, and the inner product by $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$. $\langle \mathbf{A}, \mathbf{B} \rangle = \text{Tr}(\mathbf{A}^T \mathbf{B})$. Denote the unit Euclidean ball in $\mathbb{R}^d$ by $\mathbf{B}_d$, i.e., $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq 1\}$, and the unit Euclidean sphere in $\mathbb{R}^d$ by $\mathbf{S}_{d-1}$, i.e., $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$. Further, the standard Gaussian distribution is denoted by $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The uniform distribution over the unit ball $\mathbf{B}_d$ and the unit Euclidean sphere $\mathbf{S}_{d-1}$ are denoted by $\mathcal{U}(\mathbf{B}_d)$ and $\mathcal{U}(\mathbf{S}_{d-1})$, respectively.

2. PRELIMINARIES

A. Smoothness

Definition 1. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be* $\mathbf{L_0}$***-Lipschitz*** *(also denoted by $\mathbf{f} \in \mathbf{C^{0,0}}$), if $|f(\mathbf{x}) - f(\mathbf{y})| \leq L_0 \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Definition 2. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to* ***have*** $\mathbf{L_1}$***-Lipschitz continuous gradients*** *(also called* $\mathbf{L_1}$***-smooth***, *denoted by $\mathbf{f} \in \mathbf{C^{1,1}}$), if $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_1 \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Lemma 1. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is L_1-smooth, then*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x})| \leq \frac{1}{2} L_1 \|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{1}$$

Definition 3. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to* ***have*** $\mathbf{L_2}$***-Lipschitz continuous Hessians*** *(denoted by $\mathbf{f} \in \mathbf{C^{2,2}}$), if $\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\| \leq L_2 \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Lemma 2 (Lemma 1 in [4])**.** *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is $C^{2,2}$, then*

$$\left\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \langle \nabla^2(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\right\| \leq \frac{1}{2} L_2 \|\mathbf{y} - \mathbf{x}\|^2, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{2}$$

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{1}{2} \langle \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{1}{6} L_2 \|\mathbf{y} - \mathbf{x}\|^3, \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{3}$$

B. Moments in high dimensional space

Lemma 3 (Lemma 1 in [1])**.** (a) *If M_p is defined as $M_p = \frac{1}{(2\pi)^{\frac{d}{2}}} \int \|\mathbf{u}\|^p e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u}$, then for $p = 0, 1, 2$, we have*

$$M_p \leq d^{\frac{p}{2}}. \tag{4}$$

If $p \geq 2$, then we have two-side bounds

$$d^{\frac{p}{2}} \leq M_p \leq (d + p)^{\frac{p}{2}}. \tag{5}$$

(b) *Let $\mathbf{I}_d$ be the identity matrix in $\mathbb{R}^{d \times d}$, then*

$$\int \mathbf{u}\mathbf{u}^T e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} = (2\pi)^{\frac{d}{2}} \mathbf{I}_d. \tag{6}$$

Lemma 4 (Lemma 7.3 in [5]). *Let α_d be the volume of the unit Euclidean ball $\mathbf{B}_d$ in $\mathbb{R}^d$, and β_d be the surface area of the unit Euclidean sphere $\mathbf{S}_{d-1}$ in $\mathbb{R}^d$.*

(a) *If M_p is defined as $M_p = \frac{1}{\alpha_d}\int_{\mathbf{B}_d} \|\mathbf{u}\|^p \, d\mathbf{u}$, then*

$$M_p = \frac{d}{d+p}. \tag{7}$$

(b) *Let $\mathbf{I}_d$ be the identity matrix in $\mathbb{R}^{d\times d}$, then*

$$\int_{\mathbf{S}_{d-1}} \mathbf{u}\mathbf{u}^T d\mathbf{u} = \frac{\beta_d}{d}\mathbf{I}_d. \tag{8}$$

C. Inequalities

Lemma 5. *For variables $\{\mathbf{x}_i\}_{i=1}^n$, we have*

$$\left\|\sum_{i=1}^n \mathbf{x}_i\right\|^2 \leq n\sum_{i=1}^n \|\mathbf{x}_i\|^2. \tag{9}$$

3. Smoothing technique

The smoothing technique is to utilize the integration operator to promote the differentiability [5]. More specifically, suppose that $\mathbf{u}$ is a random vector in $\mathbb{R}^d$ with density function ρ. A smooth approximation of $f:\ \mathbb{R}^d \to \mathbb{R}$ with the smoothing parameter μ is defined as:

$$f_\mu(\mathbf{x}) = \int f(\mathbf{x}+\mu\mathbf{u})\rho(\mathbf{u})d\mathbf{u}. \tag{10}$$

In general, f_μ has better properties than f.

Lemma 6. *If $f \in C^{0,0}$, then $f_\mu \in C^{0,0}$, and $L_0(f_\mu) \leq L_0(f)$.*

Lemma 7. *If $f \in C^{1,1}$, then $f_\mu \in C^{1,1}$, and $L_1(f_\mu) \leq L_1(f)$.*

Lemma 8. *If $f \in C^{2,2}$, then $f_\mu \in C^{2,2}$, and $L_2(f_\mu) \leq L_2(f)$.*

Lemma 9. *If f is convex, then f_μ is also convex.*

A. Gaussian smoothing

Gaussian smoothing is first utilized to analyze the zero-order optimization methods in [1]. Here, for the ease of presentation and understanding, we adopt a standard and simplified version introduced in [6], which is defined as follows [1]:

$$f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[f(\mathbf{x}+\mu\mathbf{u})] = \frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x}+\mu\mathbf{u})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}, \tag{11}$$

where $\mathcal{N}(\mathbf{0},\mathbf{I}_d)$ is the standard Gaussian distribution.

[1]For the general version of Gaussian smoothing function, interested readers can refer to [1].

Lemma 10. *The smoothing function f_μ is continuously differentiable, and for any $\mu > 0$, we have*

$$\nabla f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[\frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu}\mathbf{u}] = \frac{1}{(2\pi)^{\frac{d}{2}}}\int \frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu}\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}. \tag{12}$$

Lemma 11 (Theorem 1 in [1])**.** *If $f \in C^{0,0}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \mu L_0 d^{\frac{1}{2}}. \tag{13}$$

Lemma 12 (Theorem 1 in [1], Lemma 3 in [1])**.** *If $f \in C^{1,1}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \frac{\mu^2}{2}L_1 d. \tag{14}$$

$$\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \frac{\mu}{2}L_1(d+3)^{\frac{3}{2}}. \tag{15}$$

Lemma 13 (Theorem 1 in [1], Lemma 3 in [1])**.** *If $f \in C^{2,2}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\left|f_\mu(\mathbf{x}) - f(\mathbf{x}) - \frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x}), \mathbf{I}_d\rangle\right| \leq \frac{\mu^3}{6}L_2(d+3)^{\frac{3}{2}}. \tag{16}$$

$$\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \frac{\mu^2}{6}L_2(d+4)^2. \tag{17}$$

B. Uniform smoothing

Compared with Gaussian smoothing technique, uniform smoothing technique is adopted in more works [2], [3], [5], [7], [8]. Contrast with Gaussian smoothing technique, uniform smoothing technique has the following advantages: (1) the most problems are defined in a bounded space (constrainted optimization problem) rather than the whole real space required for Gaussian distribution, (2) the bounds derived from uniform smoothing scheme are slightly sharper (up to some constant factor) than that of Gaussian smoothing scheme, uniform distribution is more practical and more common [3], [5]. Following these works, the smoothed version of function f defined on uniform distribution is defined as follows:

$$f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{B}_d)}[f(\mathbf{x}+\mu\mathbf{u})] = \frac{1}{\alpha_d}\int_{\mathbf{B}_d} f(\mathbf{x}+\mu\mathbf{u})d\mathbf{u}, \tag{18}$$

where $\mu > 0, \mathcal{U}(\mathbf{B}_d)$ is the uniform distribution over the unit ball $\mathbf{B}_d$, and α_d is the volume of $\mathbf{B}_d$, i.e., $\alpha_d = \int_{\mathbf{B}_d} d\mathbf{u}$.

Lemma 14 (Lemma 2.1 in [8])**.** *The smoothing function f_μ is continuously differentiable, and for any $\mu > 0$, we have*

$$\nabla f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}]. \tag{19}$$

Lemma 15 (Lemma 4.3 in [7])**.** *If $f \in C^{0,0}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \mu L_0. \tag{20}$$

Lemma 16 (Lemma 4.1 in [5])**.** *If $f \in C^{1,1}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$|f_\mu(\mathbf{x}) - f(\mathbf{x})| \leq \frac{\mu^2 L_1}{2}. \tag{21}$$

$$\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| \leq \frac{\mu d L_1}{2}. \tag{22}$$

Lemma 17. *If $f \in C^{2,2}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$TBD \tag{23}$$

C. Coordinate uniform smoothing

The coordinate uniform smoothing is defined as follows: given a function $f : \mathbb{R}^d \to \mathbb{R}$ and a predefined smoothing/approximation parameter vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_d)^T \in \mathbb{R}^d_{++}$, we define the smoothing function of f with respect to the ith coordinate as [9]

$$f_{\mu_i}(\mathbf{x}) = \mathbf{E}_{u \sim \mathcal{U}_{[-\mu_i, \mu_i]}} f(\mathbf{x} + u\mathbf{e}_i) = \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} f(\mathbf{x} + u\mathbf{e}_i) du = \frac{1}{2} \int_{-1}^{1} f(\mathbf{x} + \mu_i v \mathbf{e}_i) dv, \tag{24}$$

where $\mathcal{U}_{[-\mu_i, \mu_i]}$ is the uniform distribution over the range $[-\mu_i, \mu_i]$, and $\mathbf{e}_i$ is a vector with i-th element being 1 and other elements being 0.

Lemma 18 (Lemma 6 in [9])**.** *The smoothing function f_{μ_i} is continuously differentiable, and for any $\mathbf{x}$, we have*

$$\frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i} = \frac{f(\mathbf{x} + \mu_i \mathbf{e}_i) - f(\mathbf{x} - \mu_i e_i)}{2\mu_i}, \tag{25}$$

where $\partial f / \partial x_i$ denotes the partial derivative with respect to x_i.

Lemma 19. *If $f \in C^{0,0}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$ and any $i \in \{1, 2, , \cdots, d\}$, we have*

$$|f_{\mu_i}(\mathbf{x}) - f(\mathbf{x})| \leq \mu_i L_0 \ (or \ \frac{\mu_i L_0}{2}). \tag{26}$$

Lemma 20 (Lemma 6 in [9])**.** *If $f \in C^{1,1}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$ and any $i \in \{1, 2, , \cdots, d\}$, we have*

$$|f_{\mu_i}(\mathbf{x}) - f(\mathbf{x})| \leq \frac{\mu_i^2 L_1}{2} \ (or \ \frac{\mu_i^2 L_1}{6}). \tag{27}$$

$$\left| \frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i} - \frac{\partial f(\mathbf{x})}{\partial x_i} \right| \leq \frac{\mu_i L_1}{2}. \tag{28}$$

Lemma 21. *If $f \in C^{2,2}: \mathbb{R}^d \to \mathbb{R}$, then for any $\mathbf{x} \in \mathbb{R}^d$ and any $i \in \{1, 2, , \cdots, d\}$, we have*

$$TBD \tag{29}$$

4. Zero-order Oracles

A. One-point zero-order oracles [8]

a) :

$$G^1_f(\mathbf{x}; \mu, \mathbf{u}) = \frac{f(\mathbf{x} + \mu \mathbf{u})}{\mu} \mathbf{u}, \tag{30}$$

where $\mu > 0$ is a smoothing parameter, and $\mathbf{u}$ is random vector drawn from a standard Gaussian distribution, i.e. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

b) :

$$G_f^1(\mathbf{x};\mu,\mathbf{u}) = d\frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu}\mathbf{u}, \tag{31}$$

where d is the dimension of the optimization variables, $\mu > 0$ is a smoothing parameter, and $\mathbf{u}$ is random unit vector drawn from a uniform distribution over a unit Euclidean sphere $\mathbf{S}_{d-1}$.

B. Two-point zero-order oracles

1) :

a) :

$$G_f^2(\mathbf{x};\mu,\mathbf{u}) = \frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x}-\mu\mathbf{u})}{2\mu}\mathbf{u}, \text{ or } G_f^2(\mathbf{x};\mu,\mathbf{u}) = \frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})}{\mu}\mathbf{u}, \tag{32}$$

where $\mu > 0$ is a smoothing parameter, and $\mathbf{u}$ is random vector drawn from a standard Gaussian distribution, i.e. $\mathcal{N}(\mathbf{0},\mathbf{I}_d)$.

b) :

$$G_f^2(\mathbf{x};\mu,\mathbf{u}) = d\frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x}-\mu\mathbf{u})}{2\mu}\mathbf{u}, \text{ or } G_f^2(\mathbf{x};\mu,\mathbf{u}) = d\frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})}{\mu}\mathbf{u}, \tag{33}$$

where d is the dimension of the optimization variables, $\mu > 0$ is a smoothing parameter, and $\mathbf{u}$ is random unit vector drawn from a uniform distribution over a unit Euclidean sphere $\mathbf{S}_{d-1}$. This zero-order oracle/gradient estimator is called RandGradEst (random gradient estimator) in [2].

2) :

$$\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q) = \frac{d}{q}\sum_{i=1}^q \frac{f(\mathbf{x}+\mu\mathbf{u}_i)-f(\mathbf{x}-\mu\mathbf{u}_i)}{2\mu}\mathbf{u}_i, \text{ or } \bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q) = \frac{d}{q}\sum_{i=1}^q \frac{f(\mathbf{x}+\mu\mathbf{u}_i)-f(\mathbf{x})}{\mu}\mathbf{u}_i, \tag{34}$$

where d is the dimension of the optimization variables, $\mu > 0$ is a smoothing parameter, and $\{\mathbf{u}\}_{i=1}^q$ are i.i.d. random unit vector drawn from a uniform distribution over a unit Euclidean sphere $\mathbf{S}_{d-1}$. Note that the smoothing parameter μ can be generalized to μ_i for $i = 1, 2, \cdots, q$. This zero-order oracle/gradient estimator is called Avg-RandGradEst (average random gradient estimator) in [2].

3) :

$$G_f^{2d}(\mathbf{x};\boldsymbol{\mu}) = \sum_{i=1}^d \frac{f(\mathbf{x}+\mu_i\mathbf{e}_i)-f(\mathbf{x}-\mu_i\mathbf{e}_i)}{2\mu}\mathbf{e}_i, \tag{35}$$

where $\mathbf{e}_i$ is a vector with i-th element being 1 and other elements being 0. Note that the smoothing parameter μ can be specialized as $\mu = \mu_i$ for $i = 1, 2, \cdots, d$. This zero-order oracle/gradient estimator is called CoordGradEst (coordinate-wise gradient estimator) in [2].

C. Properties of oracles

One-point zero-order oracles: TBD.

Two-point zero-order oracles:

1) $G_f^2(\mathbf{x};\mu,\mathbf{u})$*(RandGradEst):*

a) :

Lemma 22 ([1]). *For any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[G_f^2(\mathbf{x};\mu,\mathbf{u})\right] = \nabla f_\mu(\mathbf{x}), \tag{36}$$

where $f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}[f(\mathbf{x}+\mu\mathbf{u})]$.

Lemma 23 (Theorem 4 in [1]). *If $f \in C^{0,0}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] \le (d+4)^2 L_0^2. \tag{37}$$

Lemma 24 (Theorem 4 in [1]). *If $f \in C^{1,1}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] \le 2(d+4)\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{\mu^2}{2}L_1^2(d+6)^3. \tag{38}$$

Lemma 25 (Theorem 4 in [1]). *If $f \in C^{2,2}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] \le 2(d+4)\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{\mu^4}{18}L_2^2(d+8)^4, \tag{39}$$

where $G_f^2(\mathbf{x};\mu,\mathbf{u}) = \frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x}-\mu\mathbf{u})}{2\mu}\mathbf{u}$.

b) :

Lemma 26 (Lemma 10 in [10]). *For any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[G_f^2(\mathbf{x};\mu,\mathbf{u})\right] = \nabla f_\mu(\mathbf{x}), \tag{40}$$

where $f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{B}_d)}[f(\mathbf{x}+\mu\mathbf{u})]$.

Lemma 27. *If $f \in C^{0,0}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] \le d^2 L_0^2. \tag{41}$$

Lemma 28. *If $f \in C^{1,1}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] \le 2d\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{\mu^2 d^2 L_1^2}{2}. \tag{42}$$

Lemma 29. *If $f \in C^{2,2}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$TBD. \tag{43}$$

2) $\bar{G}_f^2(\mathbf{x};\mu,\mathbf{u})$(Avg-RandGradEst):

Lemma 30. *For any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q)\right] = \nabla f_\mu(\mathbf{x}), \tag{44}$$

where $f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{B}_d)}[f(\mathbf{x}+\mu\mathbf{u})]$.

Lemma 31. *If $f \in C^{0,0}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q)\right\|^2\right] \leq \frac{d^2L_0^2}{q}. \tag{45}$$

Lemma 32. *If $f \in C^{1,1}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q)\right\|^2\right] \leq 2\left(1+\frac{d}{q}\right)\|\nabla f(\mathbf{x})\|^2 + \left(1+\frac{1}{q}\right)\frac{\mu^2d^2L_1^2}{2}. \tag{46}$$

Lemma 33. *If $f \in C^{2,2}$, then for any $\mu > 0$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$TBD. \tag{47}$$

3) $G_f^{2d}(\mathbf{x};\mu)$(CoordGradEst):

Lemma 34. *For any $\boldsymbol{\mu} \in \mathbb{R}_{++}^d$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$G_f^{2d}(\mathbf{x};\boldsymbol{\mu}) = \sum_{i=1}^d \frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i}\mathbf{e}_i, \tag{48}$$

where $f_{\mu_i}(\mathbf{x}) = \mathbf{E}_{u\sim\mathcal{U}_{[-\mu_i,\mu_i]}} f(\mathbf{x}+u\mathbf{e}_i)$.

Lemma 35. *If $f \in C^{0,0}$, then for any $\boldsymbol{\mu} \in \mathbb{R}_{++}^d$ and $\mathbf{x} \in \mathbb{R}^d$,*

TBD

Lemma 36. *If $f \in C^{1,1}$, then for any $\boldsymbol{\mu} \in \mathbb{R}_{++}^d$ and $\mathbf{x} \in \mathbb{R}^d$,*

$$\left\|G_f^{2d}(\mathbf{x};\boldsymbol{\mu}) - \nabla f(\mathbf{x})\right\|^2 \leq \frac{dL_1^2}{4}\sum_{i=1}^d \mu_i^2. \tag{49}$$

In the special case where $\mu_i = \mu, \forall i = 1, 2, \cdots, d$, it is reduced to

$$\left\|G_f^{2d}(\mathbf{x};\boldsymbol{\mu}) - \nabla f(\mathbf{x})\right\|^2 \leq \frac{\mu^2d^2L_1^2}{4}. \tag{50}$$

5. Conclusion

Zero-order oracles are the biased estimate of the gradient of the objective function, but they are the unbiased estimate of the gradient of the smoothed objective function. Through smoothed function, zero-order oracles are associated with the true gradient and bounded by it.

6. APPENDIX

Proof of Lemma 1

Proof: Let $g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, then we have $\int_0^1 \nabla g(t)dt = g(1) - g(0) = f(\mathbf{y}) - f(\mathbf{x})$. Then,

$$\begin{aligned}
&\left|f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})\right| \\
&= \left|\int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^T(\mathbf{y} - \mathbf{x})dt - \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})\right| \\
&= \left|\int_0^1 (\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^T(\mathbf{y} - \mathbf{x})dt\right| \\
&\leq \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| dt \\
&\leq L_1 \|\mathbf{y} - \mathbf{x}\|^2 \int_0^1 tdt \\
&= \frac{1}{2} L_1 \|\mathbf{y} - \mathbf{x}\|^2,
\end{aligned} \tag{51}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality follows from the definition of L_1-smooth. ■

Proof of Lemma 2 [4]

Proof: Let $g(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$, then we have $\int_0^1 \nabla g(t)dt = g(1) - g(0) = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x})$. Then,

$$\begin{aligned}
&\left\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \langle \nabla^2(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle\right\| \\
&= \left\|\int_0^1 \langle \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x}\rangle dt - \langle \nabla^2(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle\right\| \\
&= \left\|\int_0^1 \langle \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla^2 f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle dt\right\| \\
&\leq \int_0^1 \left\|\nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla^2 f(\mathbf{x})\right\| \|\mathbf{y} - \mathbf{x}\| dt \\
&\leq L_2 \|\mathbf{y} - \mathbf{x}\|^2 \int_0^1 tdt \\
&= \frac{1}{2} L_2 \|\mathbf{y} - \mathbf{x}\|^2,
\end{aligned} \tag{52}$$

where the first inequality follows from the Cauchy-Schwartz inequality and the second inequality follows from the definition.

Let $h(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - t\langle \nabla f(x), \mathbf{y} - \mathbf{x}\rangle - \frac{t^2}{2}\langle \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$, then we have $\int_0^1 \nabla h(t)dt =$

$$
\begin{aligned}
h(1)-h(0) &= f(\mathbf{y})-f(\mathbf{x})-\langle\nabla f(x),\mathbf{y}-\mathbf{x}\rangle-\tfrac{1}{2}\langle\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle. \\
&\left|f(\mathbf{y})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle-\frac{1}{2}\langle\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle\right| \\
&= \left|\int_0^1 (\langle\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})),\mathbf{y}-\mathbf{x}\rangle-\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle-t\langle\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle)dt\right| \\
&= \left|\int_0^1 (\langle\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x})-t\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle dt\right| \\
&\leq \int_0^1 \left\|\langle\nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))-\nabla f(\mathbf{x})-t\nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})\right\| \|\mathbf{y}-\mathbf{x}\| dt \\
&\overset{(52)}{\leq} \frac{1}{2}L_2\|\mathbf{y}-\mathbf{x}\|^3\int_0^1 t^2 dt \\
&= \frac{1}{6}L_2\|\mathbf{y}-\mathbf{x}\|^3
\end{aligned}
\tag{53}
$$

■

Proof of Lemma 3 [1]

Proof: For (a), please refer to the Appendix of [1].

For (b), let $\mathbf{U}=\mathbf{u}\mathbf{u}^T$. Then, $U_{ij}=u_iu_j$. Since $\mathbf{u}$ is a random standard Gaussian vector, $u_1,u_2,\cdots,u_d$ are iid random standard Gaussian scalar. Thus, (1) $\int u_i e^{-\frac{1}{2}u_i^2}du_i=0$ (zero mean), (2) $\frac{1}{\sqrt{2\pi}}\int u_i^2 e^{-\frac{1}{2}u_i^2}du_i=1\Rightarrow\int u_i^2 e^{-\frac{1}{2}u_i^2}du_i=\sqrt{2\pi}$, (3) $\frac{1}{\sqrt{2\pi}}\int e^{-\frac{1}{2}u_i^2}du_i=1\Rightarrow\int e^{-\frac{1}{2}u_i^2}du_i=\sqrt{2\pi}$. Then, we have

$$
\begin{aligned}
\int U_{ij}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} &= \int u_iu_je^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&= \int e^{-\frac{1}{2}u_1^2}du_1\cdots\int u_ie^{-\frac{1}{2}u_i^2}du_i\cdots\int u_je^{-\frac{1}{2}u_j^2}du_j\cdots\int e^{-\frac{1}{2}u_d^2}du_d \\
&= 0.
\end{aligned}
\tag{54}
$$

$$
\begin{aligned}
\int U_{ii}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} &= \int u_i^2e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&= \int e^{-\frac{1}{2}u_1^2}du_1\cdots\int u_i^2e^{-\frac{1}{2}u_i^2}du_i\cdots\int e^{-\frac{1}{2}u_d^2}du_d \\
&= (2\pi)^{\frac{d}{2}}.
\end{aligned}
\tag{55}
$$

In conclusion,

$$
\int \mathbf{u}\mathbf{u}^Te^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}=(2\pi)^{\frac{d}{2}}\mathbf{I}_d. \tag{56}
$$

■

Proof of Lemma 4 [5]

Proof: For (a), we can directly compute M_p by using the polar coordinates,

$$
M_p=\frac{1}{\alpha_d}\int_{\mathbf{B}_d}\|\mathbf{u}\|^p d\mathbf{u}=\frac{1}{\alpha_d}\int_0^1\int_{\mathbf{S}_{d-1}} r^pr^{d-1}drd\theta=\frac{1}{\alpha_d}\int_0^1 r^pr^{d-1}dr\int_{\mathbf{S}_{d-1}}d\theta\frac{1}{d+p}\frac{\beta_p}{\alpha_p}=\frac{d}{d+p}.^2 \tag{57}
$$

[2]Section 1.2.2 in https://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/chap1-high-dim-space.pdf. See also https://en.wikipedia.org/wiki/N-sphere.

For (b), let $\mathbf{U} = \mathbf{u}\mathbf{u}^T$. Then, $U_{ij} = u_i u_j$. Therefore, if $i \neq j$, by the symmetry of the unit sphere $\mathbf{S}_{d-1}$, i.e., if $\mathbf{u} \in \mathbf{S}_{d-1}$, $\mathbf{u} = (u_1, u_2, \cdots, u_n)$, then $\mathbf{v} \in \mathbf{S}_{d-1}$ for all $\mathbf{v} = (\pm u_1, \pm u_2, \cdots, \pm u_n)$, we have

$$\int_{\mathbf{S}_{d-1}} U_{ij} d\mathbf{u} = \int_{\mathbf{S}_{d-1}} u_i u_j d\mathbf{u} = \int_{\mathbf{S}_{d-1}} -u_i u_j d\mathbf{u} = \int_{\mathbf{S}_{d-1}} -U_{ij} d\mathbf{u}. \tag{58}$$

Thus, we obtain $\int_{\mathbf{S}_{d-1}} U_{ij} d\mathbf{u} = 0$.

If $i = j$, then $U_{ii} = u_i^2$. Since

$$\int_{\mathbf{S}_{d-1}} (u_1^2 + u_2^2 + \cdots + u_d^2) d\mathbf{u} = \int_{\mathbf{S}_{d-1}} \|\mathbf{u}\|^2 d\mathbf{u} \overset{\|\mathbf{u}\|=1}{=} \int_{\mathbf{S}_{d-1}} d\mathbf{u} = \beta_d. \tag{59}$$

by symmetry, we have

$$\int_{\mathbf{S}_{d-1}} u_1^2 d\mathbf{u} = \int_{\mathbf{S}_{d-1}} u_2^2 d\mathbf{u} = \cdots = \int_{\mathbf{S}_{d-1}} u_d^2 d\mathbf{u} = \frac{\beta_d}{d}. \tag{60}$$

In conclusion, $\int_{\mathbf{S}_{d-1}} \mathbf{u}\mathbf{u}^T d\mathbf{u} = \frac{\beta_d}{d} \mathbf{I}_d$. ■

Proof of Lemma 5

Proof: Since $f(\mathbf{x}) = \|\mathbf{x}\|^2$ is convex, $f(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i) \leq \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$, i.e., $\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i\|^2$. Thus, $\left\|\sum_{i=1}^n \mathbf{x}_i\right\|^2 \leq n\sum_{i=1}^n \|\mathbf{x}_i\|^2$. ■

Proof of Lemma 6

Proof:

$$\begin{aligned} |f_\mu(\mathbf{x}) - f_\mu(\mathbf{y})| &= \left| \int [f(\mathbf{x} + \mu\mathbf{u}) - f(\mathbf{y} + \mu\mathbf{u})] \rho(\mathbf{u}) d\mathbf{u} \right| \\ &\leq \int |f(\mathbf{x} + \mu\mathbf{u}) - f(\mathbf{y} + \mu\mathbf{u})| \rho(\mathbf{u}) d\mathbf{u} \\ &\leq L_0 \|\mathbf{x} - \mathbf{y}\| \int \rho(\mathbf{u}) d\mathbf{u} \\ &= L_0 \|\mathbf{x} - \mathbf{y}\|. \end{aligned} \tag{61}$$

In conclusion, $f_\mu \in C^{0,0}$. ■

Proof of Lemma 7

Proof:

$$\begin{aligned} \|\nabla f_\mu(\mathbf{x}) - \nabla f_\mu(\mathbf{y})\| &= \left\| \int [\nabla f(\mathbf{x} + \mu\mathbf{u}) - \nabla f(\mathbf{y} + \mu\mathbf{u})] \rho(\mathbf{u}) d\mathbf{u} \right\| \\ &\leq \int \|\nabla f(\mathbf{x} + \mu\mathbf{u}) - \nabla f(\mathbf{y} + \mu\mathbf{u})\| \rho(\mathbf{u}) d\mathbf{u} \\ &\leq L_1 \|\mathbf{x} - \mathbf{y}\| \int \rho(\mathbf{u}) d\mathbf{u} \\ &= L_1 \|\mathbf{x} - \mathbf{y}\|. \end{aligned} \tag{62}$$

In conclusion, $f_\mu \in C^{1,1}$. ■

Proof of Lemma 8

Proof:

$$\begin{aligned}\left\|\nabla^2 f_\mu(\mathbf{x}) - \nabla^2 f_\mu(\mathbf{y})\right\| &= \left\| \int \left[\nabla^2 f(\mathbf{x}+\mu\mathbf{u}) - \nabla^2 f(\mathbf{y}+\mu\mathbf{u})\right] \rho(\mathbf{u}) d\mathbf{u} \right\| \\ &\leq \int \left\|\nabla^2 f(\mathbf{x}+\mu\mathbf{u}) - \nabla^2 f(\mathbf{y}+\mu\mathbf{u})\right\| \rho(\mathbf{u}) d\mathbf{u} \\ &\leq L_2 \left\|\mathbf{x}-\mathbf{y}\right\| \int \rho(\mathbf{u}) d\mathbf{u} \\ &= L_2 \left\|\mathbf{x}-\mathbf{y}\right\|. \end{aligned} \tag{63}$$

In conclusion, $f_\mu \in C^{2,2}$. ■

Proof of Lemma 9

Proof: For $\forall \theta \in [0,1]$, we have

$$\begin{aligned}\theta f_\mu(\mathbf{x}) + (1-\theta) f_\mu(\mathbf{y}) &= \theta \int f(\mathbf{x}+\mu\mathbf{u})\rho(\mathbf{u})d\mathbf{u} + (1-\theta)\int f(\mathbf{y}+\mu\mathbf{u})\rho(\mathbf{u})d\mathbf{u} \\ &= \int \left[\theta f(\mathbf{x}+\mu\mathbf{u}) + (1-\theta) f(\mathbf{y}+\mu\mathbf{u})\right]\rho(\mathbf{u})d\mathbf{u} \\ &\geq \int f\left[\theta(\mathbf{x}+\mu\mathbf{u}) + (1-\theta)(\mathbf{y}+\mu\mathbf{u})\right]\rho(\mathbf{u})d\mathbf{u} \\ &= \int f\left[\theta\mathbf{x} + (1-\theta)\mathbf{y} + \mu\mathbf{u}\right]\rho(\mathbf{u})d\mathbf{u} \\ &= f_\mu\left[\theta\mathbf{x} + (1-\theta)\mathbf{y}\right], \end{aligned} \tag{64}$$

the first inequality is due to the convexity. ■

Proof of Lemma 10 [1]

Proof: Rewrite (11) in another form by introducing a new integration variable $\mathbf{y} = \mathbf{x} + \mu\mathbf{u}$:

$$f_\mu(\mathbf{x}) = \frac{1}{\mu^d (2\pi)^{\frac{d}{2}}} \int f(\mathbf{y}) e^{-\frac{1}{2\mu^2}\|\mathbf{y}-\mathbf{x}\|^2} d\mathbf{y}. \tag{65}$$

Then,

$$\begin{aligned}\nabla f_\mu(\mathbf{x}) &= \frac{1}{\mu^{d+2}(2\pi)^{\frac{d}{2}}} \int f(\mathbf{y}) e^{-\frac{1}{2\mu^2}\|\mathbf{y}-\mathbf{x}\|^2} (\mathbf{y}-\mathbf{x}) d\mathbf{y} \\ &= \frac{1}{\mu(2\pi)^{\frac{d}{2}}} \int f(\mathbf{x}+\mu\mathbf{u}) e^{-\frac{1}{2}\|\mathbf{u}\|^2} \mathbf{u} d\mathbf{u} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int \left[\frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu}\mathbf{u}\right] e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u}. \end{aligned} \tag{66}$$

■

Proof of Lemma 11 [1]

Proof: Since $\frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} = f(\mathbf{x})$, we have

$$\begin{aligned}
|f_\mu(\mathbf{x}) - f(\mathbf{x})| &= \left|\frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x}+\mu\mathbf{u})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} - f(\mathbf{x})\right| \\
&= \left|\frac{1}{(2\pi)^{\frac{d}{2}}}\int \left[f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x})\right]e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right| \\
&\leq \frac{1}{(2\pi)^{\frac{d}{2}}}\int \left|f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x})\right|e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\leq \mu L_0 \times \frac{1}{(2\pi)^{\frac{d}{2}}}\int \|\mathbf{u}\|\, e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\overset{(4)}{\leq} \mu L_0 d^{\frac{1}{2}}.
\end{aligned} \tag{67}$$

■

Proof of Lemma 12 [1]

Proof: Since $\frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} = f(\mathbf{x})$ and $\frac{1}{(2\pi)^{\frac{d}{2}}}\int \nabla f(\mathbf{x})^T\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} = 0$, we have

$$\begin{aligned}
|f_\mu(\mathbf{x}) - f(\mathbf{x})| &= \left|\frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x}+\mu\mathbf{u})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} - f(\mathbf{x})\right| \\
&= \left|\frac{1}{(2\pi)^{\frac{d}{2}}}\int \left[f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu\nabla f(\mathbf{x})^T\mathbf{u}\right]e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right| \\
&\leq \frac{1}{(2\pi)^{\frac{d}{2}}}\int \left|f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu\nabla f(\mathbf{x})^T\mathbf{u}\right|e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\overset{(1)}{\leq} \frac{\mu^2}{2}L_1 \times \frac{1}{(2\pi)^{\frac{d}{2}}}\int \|\mathbf{u}\|^2\, e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\overset{(4)}{\leq} \frac{\mu^2}{2}L_1 d.
\end{aligned} \tag{68}$$

Further, since $\int\langle\nabla f(\mathbf{x}),\mathbf{u}\rangle\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} = \nabla f(\mathbf{x})$, we have

$$\begin{aligned}
\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| &= \left\|\frac{1}{(2\pi)^{\frac{d}{2}}}\int \frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu}\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} - \nabla f(\mathbf{x})\right\| \\
&= \left\|\frac{1}{(2\pi)^{\frac{d}{2}}}\left[\frac{1}{\mu}\int f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} - \frac{1}{\mu}\int f(\mathbf{x})\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} - \frac{1}{\mu}\int\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle\mathbf{u}e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right]\right\| \\
&\leq \frac{1}{\mu}\frac{1}{(2\pi)^{\frac{d}{2}}}\int \left|f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle\right|\|\mathbf{u}\|\, e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\overset{(1)}{\leq} \frac{1}{\mu}\frac{\mu^2 L_1}{2}\int \|\mathbf{u}\|^3\, e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u} \\
&\overset{(5)}{\leq} \frac{\mu}{2}L_1(d+3)^{\frac{3}{2}}.
\end{aligned} \tag{69}$$

■

Proof of Lemma 13 [1]

Proof: $\frac{1}{(2\pi)^{\frac{d}{2}}}\int f(\mathbf{x})e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}=f(\mathbf{x})$, $\frac{1}{(2\pi)^{\frac{d}{2}}}\int\langle\nabla f(\mathbf{x}),\mathbf{u}\rangle e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}=0$, and

$$\begin{aligned}\frac{1}{(2\pi)^{\frac{d}{2}}}\int\langle\nabla^2 f(\mathbf{x})\mathbf{u},\mathbf{u}\rangle e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}&=\frac{1}{(2\pi)^{\frac{d}{2}}}\int\sum_{i=1}^{d}\sum_{j=1}^{d}u_iu_j\nabla^2 f(\mathbf{x})_{i,j}^T e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\&=\sum_{i=1}^{d}\sum_{j=1}^{d}\left[\frac{1}{(2\pi)^{\frac{d}{2}}}\int u_iu_je^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right]\nabla^2 f(\mathbf{x})_{i,j}^T\\&\overset{(54)(55)}{=}\sum_{i=1}^{d}\nabla^2 f(\mathbf{x})_{i,i}^T\\&=\langle\nabla^2 f(\mathbf{x}),\mathbf{I}_d\rangle.\end{aligned}\tag{70}$$

As a result,

$$\begin{aligned}&\frac{1}{(2\pi)^{\frac{d}{2}}}\int\left[f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\mu\langle\nabla f(\mathbf{x}),\mathbf{u}\rangle-\frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x})\mathbf{u},\mathbf{u}\rangle\right]e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\=&f_\mu(\mathbf{x})-f(\mathbf{x})-\frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x}),\mathbf{I}_d\rangle.\end{aligned}\tag{71}$$

Thus,

$$\begin{aligned}\left|f_\mu(\mathbf{x})-f(\mathbf{x})-\frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x}),\mathbf{I}_d\rangle\right|&=\left|\frac{1}{(2\pi)^{\frac{d}{2}}}\int\left[f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\mu\langle\nabla f(\mathbf{x}),\mathbf{u}\rangle-\frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x})\mathbf{u},\mathbf{u}\rangle\right]e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right|\\&\leq\frac{1}{(2\pi)^{\frac{d}{2}}}\int\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\mu\langle\nabla f(\mathbf{x}),\mathbf{u}\rangle-\frac{\mu^2}{2}\langle\nabla^2 f(\mathbf{x})\mathbf{u},\mathbf{u}\rangle\right|e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\&\overset{(3)}{\leq}\frac{\mu^3}{6}L_2\times\frac{1}{(2\pi)^{\frac{d}{2}}}\|\mathbf{u}\|^3 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\&\overset{(5)}{\leq}\frac{\mu^3}{6}L_2(d+3)^{\frac{3}{2}}.\end{aligned}\tag{72}$$

Let $\bar{\mathbf{u}} = \langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u}\rangle \mathbf{u}$. Then, $\bar{u}_k = \sum_{i=1}^{d}\sum_{j=1}^{d} \nabla^2 f(\mathbf{x})_{i,j}^T u_i u_j u_k,\ k = 1, 2, \cdots, d.$

$$
\begin{aligned}
&\frac{1}{(2\pi)^{\frac{d}{2}}} \int \bar{u}_k e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \\
=&\frac{1}{(2\pi)^{\frac{d}{2}}} \int \sum_{i=1}^{d}\sum_{j=1}^{d} \nabla^2 f(\mathbf{x})_{i,j}^T u_i u_j u_k e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \\
=&\frac{1}{(2\pi)^{\frac{d}{2}}} \int \left[\sum_{i=1,i\neq k}^{d} \sum_{j=1,j\neq k}^{d} \nabla^2 f(\mathbf{x})_{i,j}^T u_i u_j u_k + \sum_{i=1,i\neq k}^{d} \left(\nabla^2 f(\mathbf{x})_{i,k}^T + \nabla^2 f(\mathbf{x})_{k,i}^T\right) u_i u_k^2 + \nabla^2 f(\mathbf{x})_{k,k}^T u_k^3 \right] e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \\
=& \sum_{i=1,i\neq k}^{d} \sum_{j=1,j\neq k}^{d} \nabla^2 f(\mathbf{x})_{i,j}^T \frac{1}{2\pi} \int e^{\frac{1}{2}u_1^2} du_1 \cdots \underbrace{\frac{1}{2\pi} \int u_i e^{\frac{1}{2}u_i^2} du_i}_{=0} \cdots \frac{1}{2\pi} \int u_j e^{\frac{1}{2}u_j^2} du_j \cdots \frac{1}{2\pi} \int u_k e^{\frac{1}{2}u_k^2} du_k \cdots \frac{1}{2\pi} \int u_d e^{\frac{1}{2}u_d^2} du_d \\
+& \sum_{i=1,i\neq k}^{d} \left(\nabla^2 f(\mathbf{x})_{i,k}^T + \nabla^2 f(\mathbf{x})_{k,i}^T\right) \frac{1}{2\pi} \int e^{\frac{1}{2}u_1^2} du_1 \cdots \underbrace{\frac{1}{2\pi} \int u_i e^{\frac{1}{2}u_i^2} du_i}_{=0} \cdots \frac{1}{2\pi} \int u_k^2 e^{\frac{1}{2}u_k^2} du_k \cdots \frac{1}{2\pi} \int u_d e^{\frac{1}{2}u_d^2} du_d \\
+&\nabla^2 f(\mathbf{x})_{k,k}^T \frac{1}{2\pi} \int e^{\frac{1}{2}u_1^2} du_1 \cdots \underbrace{\frac{1}{2\pi} \int u_k^3 e^{\frac{1}{2}u_k^2} du_k}_{=0} \cdots \frac{1}{2\pi} \int u_d e^{\frac{1}{2}u_d^2} du_d \\
=&0.
\end{aligned}
\tag{73}
$$

Here, we use the result that the third moment of the random standard Gaussian scalar is 0 [11].

Then, we have

$$
\begin{aligned}
&\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| \\
=&\left\| \frac{1}{(2\pi)^{\frac{d}{2}}} \int \frac{f(\mathbf{x}+\mu\mathbf{u})}{\mu} \mathbf{u} e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} - \nabla f(\mathbf{x}) \right\| \\
=&\left\| \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{\mu} \int \left[f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu\langle \nabla f(\mathbf{x}), \mathbf{u}\rangle - \frac{\mu^2}{2}\langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u}\rangle \right] \mathbf{u} e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \right\| \\
\leq& \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{\mu} \int \left| f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu\langle \nabla f(\mathbf{x}), \mathbf{u}\rangle - \frac{\mu^2}{2}\langle \nabla^2 f(\mathbf{x})\mathbf{u}, \mathbf{u}\rangle \right| \|\mathbf{u}\| e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \\
\overset{(3)}{\leq}& \frac{1}{\mu} \frac{\mu^3}{6} L_2 \frac{1}{(2\pi)^{\frac{d}{2}}} \int \|\mathbf{u}\|^4 e^{-\frac{1}{2}\|\mathbf{u}\|^2} d\mathbf{u} \\
\overset{(5)}{\leq}& \frac{\mu^2}{6} L_2 (d+4)^2.
\end{aligned}
\tag{74}
$$

■

Proof of Lemma 14 [8]

Proof: Following from Stoke's theorem, we have

$$
\nabla \int_{\mu\mathbf{B}_d} f(\mathbf{x}+\mathbf{v}) d\mathbf{v} = \int_{\mu\mathbf{S}_{d-1}} f(\mathbf{x}+\mathbf{u}) \frac{\mathbf{u}}{\|\mathbf{u}\|} d\mathbf{u}, \tag{75}
$$

where $\mu\mathbf{B}_d$ and $\mu\mathbf{S}_{d-1}$ are the ball and sphere of radius μ.

By definition, $f_\mu(\mathbf{x}) = \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{B}_d)}[f(\mathbf{x}+\mu\mathbf{u})] = \frac{1}{\alpha_d}\int_{\mathbf{B}_d} f(\mathbf{x}+\mu\mathbf{u})d\mathbf{u} = \frac{1}{\int_{\mu\mathbf{B}_d} d\mathbf{v}}\int_{\mu\mathbf{B}_d} f(\mathbf{x}+\mathbf{v})d\mathbf{v} = \frac{1}{\mu^d\alpha_d}\int_{\mu\mathbf{B}_d} f(\mathbf{x}+\mathbf{v})d\mathbf{v}$.[3] Similarly, $\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}] = \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}} f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}d\mathbf{u} = \frac{1}{\int_{\mu\mathbf{S}_{d-1}} d\mathbf{u}}\int_{\mu\mathbf{S}_{d-1}} f(\mathbf{x}+\mathbf{u})\frac{\mathbf{u}}{\|\mathbf{u}\|}d\mathbf{u} = \frac{1}{\mu^{d-1}\beta_d}\int_{\mu\mathbf{S}_{d-1}} f(\mathbf{x}+\mathbf{u})\frac{\mathbf{u}}{\|\mathbf{u}\|}d\mathbf{u}$.

$$\begin{aligned}
\nabla f_\mu(\mathbf{x}) =& \frac{1}{\mu^d\alpha_d}\nabla\int_{\mu\mathbf{B}_d} f(\mathbf{x}+\mathbf{v})d\mathbf{v}\\
\overset{(75)}{=}& \frac{1}{\mu^d\alpha_d}\int_{\mu\mathbf{S}_{d-1}} f(\mathbf{x}+\mathbf{u})\frac{\mathbf{u}}{\|\mathbf{u}\|}d\mathbf{u}\\
=& \frac{\beta_d}{\mu\alpha_d}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}]\\
=& \frac{d}{\mu}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}]\\
=& \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}].
\end{aligned} \tag{76}$$

Note that the smoothed function f_μ is differentiable even when f is not. ■

Proof of Lemma 15

Proof: Indeed, for any $\mathbf{x}\in\mathbb{R}^d$, we have $f_\mu(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{\alpha_d}\int_{\mathbf{B}_d}(f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}))d\mathbf{u}$. Therefore,

$$\begin{aligned}
|f_\mu(\mathbf{x}) - f(\mathbf{x})| =& \left|\frac{1}{\alpha_d}\int_{\mathbf{B}_d}(f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}))d\mathbf{u}\right|\\
\leq& \frac{1}{\alpha_d}\int_{\mathbf{B}_d}|f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x})|\,d\mathbf{u}\\
\leq& \mu L_0\frac{1}{\alpha_d}\int_{\mathbf{B}_d}\|\mathbf{u}\|\,d\mathbf{u}\\
\overset{(7)}{\leq}& \frac{d}{d+1}\mu L_0\\
\leq& \mu L_0.
\end{aligned} \tag{77}$$

■

Proof of Lemma 16

Proof: We already know that for any $\mathbf{x}\in\mathbb{R}^d$, we have $f_\mu(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{\alpha_d}\int_{\mathbf{B}_d}(f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}))d\mathbf{u}$. Further, if f is differentiable at $\mathbf{x}$, then $f_\mu(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{\alpha_d}\int_{\mathbf{B}_d}(f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu\nabla f(\mathbf{x})^T\mathbf{u})d\mathbf{u}$. This is because that

[3]https://en.wikipedia.org/wiki/Volume_of_an_n-ball

$\int_{\mathbf{B}_d} \mathbf{u} d\mathbf{u} = \mathbf{0}$. Therefore,

$$\begin{aligned}
|f_\mu(\mathbf{x}) - f(\mathbf{x})| &= \left| \frac{1}{\alpha_d} \int_{\mathbf{B}_d} (f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu \nabla f(\mathbf{x})^T \mathbf{u}) d\mathbf{u} \right| \\
&\leq \frac{1}{\alpha_d} \int_{\mathbf{B}_d} \left| (f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \mu \nabla f(\mathbf{x})^T \mathbf{u}) \right| d\mathbf{u} \\
&\overset{(1)}{\leq} \frac{\mu^2 L_1}{2} \frac{1}{\alpha_d} \int_{\mathbf{B}_d} \|\mathbf{u}\|^2 d\mathbf{u} \\
&\overset{(7)}{\leq} \frac{\mu^2 L_1}{2} \frac{d}{d+2} \\
&\leq \frac{\mu^2 L_1}{2}.
\end{aligned} \tag{78}$$

Since 1) $\int_{\mathbf{S}_{d-1}} \mathbf{u} d\mathbf{u} = 0$, 2) $\frac{1}{\beta_d} \int_{\mathbf{S}_{d-1}} \frac{d}{\mu} \langle \nabla f(\mathbf{x}), \mu\mathbf{u} \rangle \mathbf{u} d\mathbf{u} = \frac{1}{\beta_d} \int_{\mathbf{S}_{d-1}} d\mathbf{u}\mathbf{u}^T \nabla f(\mathbf{x}) d\mathbf{u} = \frac{d}{\beta_d} \int_{\mathbf{S}_{d-1}} \mathbf{u}\mathbf{u}^T d\mathbf{u} \nabla f(\mathbf{x}) \overset{(4)}{=} \nabla f(\mathbf{x})$ (you can expand it and write it in the form of scalar/coordinate product), we have

$$\begin{aligned}
\|\nabla f_\mu(\mathbf{x}) - \nabla f(\mathbf{x})\| &= \left\| \frac{1}{\beta_d} \left[\frac{d}{\mu} \int_{\mathbf{S}_{d-1}} f(\mathbf{x}+\mu\mathbf{u}) \mathbf{u} d\mathbf{u} \right] - \nabla f(\mathbf{x}) \right\| \\
&= \left\| \frac{1}{\beta_d} \left[\frac{d}{\mu} \int_{\mathbf{S}_{d-1}} f(\mathbf{x}+\mu\mathbf{u}) \mathbf{u} d\mathbf{u} - \frac{d}{\mu} \int_{\mathbf{S}_{d-1}} f(\mathbf{x}) \mathbf{u} d\mathbf{u} - \frac{d}{\mu} \int_{\mathbf{S}_{d-1}} \langle \nabla f(\mathbf{x}), \mu\mathbf{u} \rangle \mathbf{u} d\mathbf{u} \right] \right\| \\
&\leq \frac{d}{\beta_d \mu} \int_{\mathbf{S}_{d-1}} |f(\mathbf{x}+\mu\mathbf{u}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mu\mathbf{u} \rangle| \, \|\mathbf{u}\| \, d\mathbf{u} \\
&\overset{(1)}{\leq} \frac{d}{\beta_d \mu} \frac{\mu^2 L_1}{2} \int_{\mathbf{S}_{d-1}} \|\mathbf{u}\|^3 d\mathbf{u} \\
&\overset{\|\mathbf{u}\|=1}{=} \frac{d}{\beta_d \mu} \frac{\mu^2 L_1}{2} \beta_d \\
&= \frac{\mu d L_1}{2}.
\end{aligned} \tag{79}$$

■

Proof of Lemma 18

Proof: Since $\frac{\partial f(\mathbf{x}+u\mathbf{e}_i)}{\partial x_i} = \frac{\partial f(\mathbf{x}+u\mathbf{e}_i)}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{x_i} = \nabla f(\mathbf{x}+u\mathbf{e}_i)\mathbf{e}_i$, $\frac{\partial f(\mathbf{x}+u\mathbf{e}_i)}{\partial u} = \nabla f(\mathbf{x}+u\mathbf{e}_i)\mathbf{e}_i$, $\frac{f(\mathbf{x}+u\mathbf{e}_i)}{\partial x_i} = \frac{f(\mathbf{x}+u\mathbf{e}_i)}{\partial u}$.

$$\begin{aligned}
\frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i} &= \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \frac{\partial f(\mathbf{x}+u\mathbf{e}_i)}{\partial x_i} du \\
&= \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \frac{\partial f(\mathbf{x}+u\mathbf{e}_i)}{\partial u} du \\
&= \frac{1}{2\mu_i} f(\mathbf{x}+u\mathbf{e}_i)|_{u=-\mu_i}^{u=\mu_i} \\
&= \frac{f(\mathbf{x}+\mu_i\mathbf{e}_i) - f(\mathbf{x}-\mu_i\mathbf{e}_i)}{2\mu_i}.
\end{aligned} \tag{80}$$

■

Proof of Lemma 19

Proof:

$$\begin{aligned}
|f_{\mu_i}(\mathbf{x}) - f(\mathbf{x})| &= \left| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} f(\mathbf{x} + u\mathbf{e}_i) du - f(\mathbf{x}) \right| \\
&= \left| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left[f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) \right] du \right| \\
&\leq \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left| f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) \right| du \\
&\leq \frac{L_0}{2\mu_i} \int_{-\mu_i}^{\mu_i} |u| \, du \\
&= \frac{L_0}{2\mu_i} \cdot \mu_i^2 \\
&= \frac{\mu_i L_0}{2}.
\end{aligned} \tag{81}$$

■

Proof of Lemma 20

Proof: Since $\frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} u \nabla f(\mathbf{x})^T \mathbf{e}_i = 0$, we have $f_{\mu_i}(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left[f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) - u \nabla f(\mathbf{x})^T \mathbf{e}_i \right] du$. Then,

$$\begin{aligned}
|f_{\mu_i}(\mathbf{x}) - f(\mathbf{x})| &= \left| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} f(\mathbf{x} + u\mathbf{e}_i) du - f(\mathbf{x}) \right| \\
&= \left| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left[f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) \right] du \right| \\
&= \left| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left[f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) - u \nabla f(\mathbf{x})^T \mathbf{e}_i \right] du \right| \\
&\leq \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left| f(\mathbf{x} + u\mathbf{e}_i) - f(\mathbf{x}) - u \nabla f(\mathbf{x})^T \mathbf{e}_i \right| du \\
&\overset{(1)}{\leq} \frac{1}{2\mu_i} \cdot \frac{L_1}{2} \int_{-\mu_i}^{\mu_i} u^2 du \\
&= \frac{\mu_i^2 L_1}{6}.
\end{aligned} \tag{82}$$

$$\begin{aligned}
\left\| \frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i} - \frac{\partial f(\mathbf{x})}{\partial x_i} \right\| &= \left\| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \nabla f(\mathbf{x} + u\mathbf{e}_i) \mathbf{e}_i du - \nabla f(\mathbf{x}) \mathbf{e}_i \right\| \\
&= \left\| \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left[\nabla f(\mathbf{x} + u\mathbf{e}_i) - \nabla f(\mathbf{x}) \right] \mathbf{e}_i du \right\| \\
&\leq \frac{1}{2\mu_i} \int_{-\mu_i}^{\mu_i} \left\| \nabla f(\mathbf{x} + u\mathbf{e}_i) - \nabla f(\mathbf{x}) \right\| du \\
&\leq \frac{L_1}{2\mu_i} \int_{-\mu_i}^{\mu_i} |u| du \\
&= \frac{L_1}{2\mu_i} \cdot \mu_i^2 \\
&= \frac{\mu_i L_1}{2}.
\end{aligned} \tag{83}$$

■

Proof of Lemma 23

Proof:

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{(2\pi)^{\frac{d}{2}}}\int\left\|\frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})}{\mu}\mathbf{u}\right\|^2 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &\overset{(1)}{\leq} L_0^2\frac{1}{(2\pi)^{\frac{d}{2}}}\int\|\mathbf{u}\|^4 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &\overset{(5)}{\leq} L_0^2(d+4)^2.\end{aligned} \tag{84}$$

■

Proof of Lemma 24

Proof:

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{(2\pi)^{\frac{d}{2}}}\int\left\|\frac{f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})}{\mu}\mathbf{u}\right\|^2 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\frac{1}{\mu^2}\int|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})|^2\|\mathbf{u}\|^2 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &= \frac{1}{(2\pi)^{\frac{d}{2}}}\frac{1}{\mu^2}\int|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle+\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle|^2\|\mathbf{u}\|^2 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &\overset{(9)}{\leq} \frac{1}{(2\pi)^{\frac{d}{2}}}\frac{1}{\mu^2}\int\left[2|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle|^2+2|\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle|^2\right]\|\mathbf{u}\|^2 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\\ &\overset{(1)}{\leq} \frac{1}{\mu^2}\left[2\mu^4\frac{L_1^2}{4}\frac{1}{(2\pi)^{\frac{d}{2}}}\int\|\mathbf{u}\|^6 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}+2\mu^2\|\nabla f(\mathbf{x})\|^2\frac{1}{(2\pi)^{\frac{d}{2}}}\int\|\mathbf{u}\|^4 e^{-\frac{1}{2}\|\mathbf{u}\|^2}d\mathbf{u}\right]\\ &\overset{(5)}{\leq} \frac{\mu^2}{2}L_1^2(d+6)^3+2\|\nabla f(\mathbf{x})\|^2(d+4)^2.\end{aligned} \tag{85}$$

This bound can be further strengthened to $\frac{\mu^2}{2}L_1^2(d+6)^3+2\|\nabla f(\mathbf{x})\|^2(d+4)$ in [1]. ■

Proof of Lemma 25

Proof: Please refer to Theorem 4 in [1]. ■

Proof of Lemma 26 (Lemma 10 in [10])

Proof: Since $\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\mathbf{u}=0$, $\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}f(\mathbf{x})\mathbf{u}=0$. Then, we have

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[G_f^2(\mathbf{x};\mu,\mathbf{u})] &= \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}-\frac{d}{\mu}f(\mathbf{x})\mathbf{u}\right]\\ &= \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}\right]-\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x})\mathbf{u}\right]\\ &\overset{(19)}{=} \nabla f_\mu(\mathbf{x}).\end{aligned} \tag{86}$$

Due to the symmetry, if $\mathbf{u}\in\mathbf{S}_{d-1}$, then $\mathbf{v}=-\mathbf{u}\in\mathbf{S}_{d-1}$. Thus,

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}-\mu\mathbf{u})\mathbf{u}\right]=-\mathbf{E}_{\mathbf{v}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{v})\mathbf{v}\right]=-\nabla f_\mu(\mathbf{x}). \tag{87}$$

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}[G_f^2(\mathbf{x};\mu,\mathbf{u})] &= \frac{1}{2}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}-\frac{d}{\mu}f(\mathbf{x}-\mu\mathbf{u})\mathbf{u}\right]\\ &= \frac{1}{2}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}+\mu\mathbf{u})\mathbf{u}\right]-\frac{1}{2}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\frac{d}{\mu}f(\mathbf{x}-\mu\mathbf{u})\mathbf{u}\right]\\ &\overset{(19)}{=} \frac{1}{2}\nabla f_\mu(\mathbf{x})+\frac{1}{2}\nabla f_\mu(\mathbf{x})\\ &= \nabla f_\mu(\mathbf{x}).\end{aligned} \tag{88}$$

■

Proof of Lemma 27

Proof:

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}}\frac{d^2}{\mu^2}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})\right|^2\|\mathbf{u}\|^2\,d\mathbf{u}\\ &\leq \frac{d^2}{\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}\mu^2L_0^2\|\mathbf{u}\|^4\,d\mathbf{u}\\ &\overset{\|\mathbf{u}\|=1}{=} \frac{d^2}{\beta_d\mu^2}\cdot\mu^2L_0^2\beta_d\\ &= d^2L_0^2.\end{aligned} \tag{89}$$

$$\begin{aligned}\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}}\frac{d^2}{4\mu^2}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x}-\mu\mathbf{u})\right|^2\|\mathbf{u}\|^2\,d\mathbf{u}\\ &= \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}}\frac{d^2}{4\mu^2}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})+f(\mathbf{x})-f(\mathbf{x}-\mu\mathbf{u})\right|^2\|\mathbf{u}\|^2\,d\mathbf{u}\\ &\overset{(9)}{\leq} \frac{d^2}{4\beta_d\mu^2}\left[2\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})\right|^2+2\left|f(\mathbf{x})-f(\mathbf{x}-\mu\mathbf{u})\right|^2\right]\\ &\leq \frac{d^2}{4\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}4\mu^2L_0^2\|\mathbf{u}\|^4\,d\mathbf{u}\\ &\overset{\|\mathbf{u}\|=1}{=} \frac{d^2}{4\beta_d\mu^2}\cdot4\mu^2L_0^2\beta_d\\ &= d^2L_0^2.\end{aligned} \tag{90}$$

■

Proof of Lemma 28

Proof:

$$\begin{aligned}
\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}}\frac{d^2}{\mu^2}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})\right|^2\|\mathbf{u}\|^2 d\mathbf{u}\\
&\overset{\|\mathbf{u}\|=1}{=} \frac{d^2}{\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle+\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle\right|^2 d\mathbf{u}\\
&\leq \frac{d^2}{\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}\left[2\left(f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})-\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle\right)^2+2\left(\langle\nabla f(\mathbf{x}),\mu\mathbf{u}\rangle\right)^2\right]d\mathbf{u}\\
&\overset{(1)}{\leq} \frac{d^2}{\beta_d\mu^2}\left[\int_{\mathbf{S}_{d-1}}2\left(\frac{L_1}{2}\mu^2\|\mathbf{u}\|^2\right)^2 d\mathbf{u}+2\mu^2\int_{\mathbf{S}_{d-1}}\nabla f(\mathbf{x})^T\mathbf{u}\mathbf{u}^T\nabla f(\mathbf{x})d\mathbf{u}\right]\\
&\overset{(4)}{=} \frac{d^2}{\beta_d\mu^2}\left[\frac{L_1^2\mu^4}{2}\beta_d+2\mu^2\frac{\beta_d}{d}\|\nabla f(\mathbf{x})\|^2\right]\\
&= 2d\|\nabla f(\mathbf{x})\|^2+\frac{\mu^2d^2L_1^2}{2}.
\end{aligned} \tag{91}$$

$$\begin{aligned}
\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u})\right\|^2\right] &= \frac{1}{\beta_d}\int_{\mathbf{S}_{d-1}}\frac{d^2}{4\mu^2}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x}-\mu\mathbf{u})\right|^2\|\mathbf{u}\|^2 d\mathbf{u}\\
&\overset{\|\mathbf{u}\|=1}{=} \frac{d^2}{4\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})+f(\mathbf{x})-f(\mathbf{x}-\mu\mathbf{u})\right|^2 d\mathbf{u}\\
&\leq \frac{d^2}{4\beta_d\mu^2}\int_{\mathbf{S}_{d-1}}\left[2\left|f(\mathbf{x}+\mu\mathbf{u})-f(\mathbf{x})\right|^2+2\left|f(\mathbf{x})-f(\mathbf{x}-\mu\mathbf{u})\right|^2\right]d\mathbf{u}\\
&\overset{(91)}{\leq} \frac{d^2}{4\beta_d\mu^2}\left[2\left(\frac{L_1^2\mu^4}{2}\beta_d+2\mu^2\frac{\beta_d}{d}\|\nabla f(\mathbf{x})\|^2+\frac{L_1^2\mu^4}{2}\beta_d+2\mu^2\frac{\beta_d}{d}\|\nabla f(\mathbf{x})\|^2\right)\right]\\
&= 2d\|\nabla f(\mathbf{x})\|^2+\frac{\mu^2d^2L_1^2}{2}.
\end{aligned} \tag{92}$$

■

Proof of Lemma 30

Proof: Since $\{\mathbf{u}_i\}_{i=1}^q$ are i.i.d. random vectors, we have

$$\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q)\right] = \frac{1}{q}\sum_{i=1}^q\mathbf{E}_{\mathbf{u}_i\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[G_f^2(\mathbf{x};\mu,\mathbf{u}_i)\right] = \nabla f_\mu(\mathbf{x}). \tag{93}$$

■

Proof of Lemma 32

Proof:

$$\begin{aligned}
\mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|\bar{G}_f^2(\mathbf{x};\mu,\{\mathbf{u}_i\}_{i=1}^q)\right\|^2\right] &= \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|\frac{1}{q}\sum_{i=1}^q G_f^2(\mathbf{x};\mu,\mathbf{u}_i)\right\|^2\right] \\
&= \mathbf{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|\frac{1}{q}\sum_{i=1}^q \left(G_f^2(\mathbf{x};\mu,\mathbf{u}_i)-\nabla f_\mu(\mathbf{x})\right)+\nabla f_\mu(\mathbf{x})\right\|^2\right] \\
&= \left\|\nabla f_\mu(\mathbf{x})\right\|^2 + \frac{1}{q}\mathbf{E}_{\mathbf{u}_1\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u}_1)-\nabla f_\mu(\mathbf{x})\right\|^2\right] \\
&\leq \left\|\nabla f_\mu(\mathbf{x})\right\|^2 + \frac{1}{q}\mathbf{E}_{\mathbf{u}_1\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left\|G_f^2(\mathbf{x};\mu,\mathbf{u}_1)\right\|^2\right] \\
&\overset{((22)42)}{\leq} 2\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{\mu^2 d^2 L_1^2}{2} + \frac{2d}{q}\left\|\nabla f(\mathbf{x})\right\|^2 + \frac{1}{q}\frac{\mu^2 d^2 L_1^2}{2} \\
&= 2\left(1+\frac{d}{q}\right)\left\|\nabla f(\mathbf{x})\right\|^2 + \left(1+\frac{1}{q}\right)\frac{\mu^2 d^2 L_1^2}{2}.
\end{aligned} \tag{94}$$

The third equality is due to the independence of $\mathbf{u}_i$ and

$$\begin{aligned}
\mathbf{E}_{\mathbf{u}_i\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left(G_f^2(\mathbf{x};\mu,\mathbf{u}_i)-\nabla f_\mu(\mathbf{x})\right)^T\nabla f_\mu(\mathbf{x})\right] &= \mathbf{E}_{\mathbf{u}_i\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[G_f^2(\mathbf{x};\mu,\mathbf{u}_i)^T\right]\nabla f_\mu(\mathbf{x}) - \left\|\nabla f_\mu(\mathbf{x})\right\|^2 \\
&\overset{(26)}{=} \left\|\nabla f_\mu(\mathbf{x})\right\|^2 - \left\|\nabla f_\mu(\mathbf{x})\right\|^2 \\
&= 0,
\end{aligned} \tag{95}$$

$$\mathbf{E}_{\mathbf{u}_i,\mathbf{u}_j\sim\mathcal{U}(\mathbf{S}_{d-1})}\left[\left(G_f^2(\mathbf{x};\mu,\mathbf{u}_i)-\nabla f_\mu(\mathbf{x})\right)^T\left(G_f^2(\mathbf{x};\mu,\mathbf{u}_j)-\nabla f_\mu(\mathbf{x})\right)\right] \overset{(26)}{=} 0. \tag{96}$$

■

Proof of Lemma 34

Proof:

$$G_f^{2d}(\mathbf{x};\mu) = \sum_{k=1}^d \frac{f(\mathbf{x}+\mu\mathbf{e}_k)-f(\mathbf{x}-\mu\mathbf{e}_k)}{2\mu}\mathbf{e}_k = \sum_{k=1}^d \frac{\partial f_{\mu_k}(\mathbf{x})}{\partial x_k}\mathbf{e}_k. \tag{97}$$

The second equation is due to the Lemma 18. ■

Proof of Lemma 36

Proof:

$$\begin{aligned}
\left\|G_f^{2d}(\mathbf{x};\mu)-\nabla f(\mathbf{x})\right\|^2 &= \left\|\sum_{i=1}^d\left(\frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i}-\frac{\partial f(\mathbf{x})}{\partial x_i}\right)\mathbf{e}_i\right\|^2 \\
&\overset{(9)}{\leq} d\sum_{i=1}^d\left|\frac{\partial f_{\mu_i}(\mathbf{x})}{\partial x_i}-\frac{\partial f(\mathbf{x})}{\partial x_i}\right|^2 \\
&\overset{(27)}{\leq} d\sum_{i=1}^d\frac{\mu_i^2 L_1^2}{4} \\
&= \frac{dL_1^2}{4}\sum_{i=1}^d\mu_i^2.
\end{aligned} \tag{98}$$

In the special case where $\mu_i = \mu, \forall i = 1, 2, \cdots, d$, the equation above can be reduced to

$$\left\| G_f^{2d}(\mathbf{x}; \mu) - \nabla f(\mathbf{x}) \right\|^2 \leq \frac{\mu^2 d^2 L_1^2}{4}. \tag{99}$$

■

REFERENCES

[1] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.

[2] S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini, "Zeroth-order stochastic variance reduction for nonconvex optimization," in *Advances in Neural Information Processing Systems*, pp. 3727–3737, 2018.

[3] S. Liu and P.-Y. Chen, "Zeroth-order optimization and its application to adversarial machine learning," *Intelligent Informatics*, p. 25, 2018.

[4] Y. Nesterov and B. T. Polyak, "Cubic regularization of newton method and its global performance," *Mathematical Programming*, vol. 108, no. 1, pp. 177–205, 2006.

[5] X. Gao, B. Jiang, and S. Zhang, "On the information-adaptive variants of the admm: an iteration complexity perspective," *Journal of Scientific Computing*, vol. 76, no. 1, pp. 327–363, 2018.

[6] D. Hajinezhad, M. Hong, and A. Garcia, "Zone: Zeroth order nonconvex multi-agent optimization over networks," *IEEE Transactions on Automatic Control*, 2019.

[7] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[8] A. D. Flaxman, A. T. Kalai, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: gradient descent without a gradient," in *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, Society for Industrial and Applied Mathematics, 2005.

[9] X. Lian, H. Zhang, C.-J. Hsieh, Y. Huang, and J. Liu, "A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order," in *Advances in Neural Information Processing Systems*, pp. 3054–3062, 2016.

[10] O. Shamir, "An optimal algorithm for bandit and zero-order convex optimization with two-point feedback.," *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[11] A. Winkelbauer, "Moments and absolute moments of the normal distribution," *arXiv preprint arXiv:1209.4340*, 2012.