Stochastic Zeroth Order Gradient and Hessian Estimators: Variance Reduction and Refined Bias Bounds

Yasong Feng* and Tianyu Wang†

Abstract

We study stochastic zeroth order gradient and Hessian estimators for real-valued functions in $\mathbb{R}^n$. We show that, via sampling from Stiefel's manifolds, the variance of the stochastic finite difference estimators can be significantly reduced, without losing any bias accuracy. In particular, we design estimators for smooth functions such that, if one uses k random directions sampled from the Stiefel's manifold $\mathrm{St}(n,k)$ and finite-difference granularity δ, the variance of the gradient estimator is bounded by $\mathcal{O}\left(\left(\frac{n}{k}-1\right)+\left(\frac{n^2}{k}-n\right)\delta^2+\frac{n^2\delta^4}{k}\right)$, and the variance of the Hessian estimator is bounded by $\mathcal{O}\left(\left(\frac{n^2}{k^2}-1\right)+\left(\frac{n^4}{k^2}-n^2\right)\delta^2+\frac{n^4\delta^4}{k^2}\right)$. When $k=n$, the variances become negligibly small. In addition, the bias of both gradient and Hessian estimators for smooth function f is of order $\mathcal{O}\left(\delta^2\Gamma\right)$, where δ is the finite-difference granularity, and Γ depends on high order derivatives of f. Our results are evidenced by empirical observations.

1 Introduction

Since Newton's time, people have been using finite difference principles to estimate derivatives. This classic problem has recently revived, as tasks of stochastic derivative estimation in high dimension become prevalent.

Various bias bounds have been derived for gradient and Hessian estimators (e.g., Flaxman et al., 2005; Nesterov and Spokoiny, 2017; Balasubramanian and Ghadimi, 2021; Wang, 2022). Yet the statistical convergence to these bias bounds is slow due to large variance, especially in high dimensional spaces. The bias of a gradient estimator $\widehat{\nabla} f(x)$ is

$$\left\|\mathbb{E}\left[\widehat{\nabla} f(x)\right]-\nabla f(x)\right\|,$$

and the bias of a Hessian estimator is similarly defined. In practice, the estimation error of $\widehat{\nabla} f(x)$ is measured by $\left\|\widehat{\nabla} f(x)-\nabla f(x)\right\|$ (similarly for the Hessian counterpart). Unless the estimator is highly concentrated around its expectation, the theoretical bias bound may not be aligned with the empirical observations. This discrepancy calls for careful study on the variance and variance reduction for the estimators.

To this end, we introduce variance-reduced methods for stochastic zeroth order gradient and Hessian estimation, and provide performance guarantees for these methods. For estimating the gradient of a function f in $\mathbb{R}^n$, we propose to uniformly sample a matrix $[v_1, v_2, \cdots, v_k]$ from the real Stiefel's manifold $\mathrm{St}(n,k) := \{X \in \mathbb{R}^{n\times k} : X^\top X = I\}$, and estimate the gradient of f at x by

$$\widehat{\nabla} f_k^\delta(x) := \frac{n}{2\delta k}\sum_{i=1}^{k}\left(f(x+\delta v_i)-f(x-\delta v_i)\right)v_i, \tag{1}$$

where δ is the finite difference granularity. When $k=1$, sampling is over the unit sphere and (Eq. 1) reduces to the estimator introduced by Flaxman et al. (2005).

*ysfeng20@fudan.edu.cn
†wangtianyu@fudan.edu.cn

We show that the variance of (Eq. 1) for a $(3, L_3)$-smooth (See Definition 1) function f satisfies

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right]\right\|^2\right] \\ \leq\ &\left(\frac{n}{k} - 1\right)\|\nabla f(x)\|^2 + \frac{L_3\delta^2}{3}\left(\frac{n^2}{k} - n\right)\|\nabla f(x)\| + \frac{L_3^2 n^2 \delta^4}{36k}, \quad \forall x \in \mathbb{R}^n,\end{aligned}$$

where $\|\cdot\|$ is the Euclidean norm. When $k = n$, the variance of the estimator becomes negligibly small.

For estimating the Hessian of a function f in $\mathbb{R}^n$, we propose to independently uniformly sample two matrices $[v_1, v_2, \cdots, v_k]$ and $[w_1, w_2, \cdots, w_k]$ from the Stiefel's manifold $\mathrm{St}(n, k)$ and estimate the Hessian of f at x by

$$\begin{aligned}\widehat{\mathrm{H}} f_k^\delta(x) := \frac{n^2}{8\delta^2 k^2}\sum_{i,j=1}^{k}\ &(f(x + \delta v_i + \delta w_j) - f(x - \delta v_i + \delta w_j) - f(x + \delta v_i - \delta w_j) + f(x - \delta v_i - \delta w_j)) \\ &\cdot (v_i w_j^\top + w_j v_i^\top),\end{aligned} \tag{2}$$

where δ is the finite difference step size. When $k = 1$, the sampling is over the unit sphere and (Eq. 2) reduces to the one introduced by Wang (2022).

The variance of (Eq. 2) for a $(4, L_4)$-smooth and $(6, L_6)$-smooth (See Definition 1) function f satisfies, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}&\mathbb{E}\left[\left\|\widehat{\mathrm{H}} f_k^\delta(x) - \mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(x)\right]\right\|_F^2\right] \\ \leq\ &\left\|\nabla^2 f(x)\right\|_F^2\left(\frac{n^2}{k^2} - 1\right) + 2\delta^2 L_4\left\|\nabla^2 f(x)\right\|\left(\frac{n^4}{k^2} - n^2\right) + \mathcal{O}\left(\left(L_6 n^2\|\nabla^2 f(0)\| + \frac{n^4 L_4^2}{k^2}\right)\delta^4\right),\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|$ is the spectral norm. Similar to the gradient case, the variance of (Eq. 2) becomes negligibly small when $k = n$.

In addition, the above estimators do not sacrifice any bias accuracy. The bias of (Eq. 1) for sufficently smooth f at x is of order

$$\mathcal{O}\left(\delta^2 \Gamma_3(x)\right),$$

where $\Gamma_3(x)$ depends on the third order derivatives of f at x. Similar results hold for the Hessian estimator. The bias of (Eq. 2) for sufficiently smooth f at x is of order

$$\mathcal{O}\left(\delta^2 \Gamma_4(x)\right),$$

where $\Gamma_4(x)$ depends on the fourth order derivatives of f at x. These refined bias bounds improve best previous results on bias of the estimators (Flaxman et al., 2005; Wang, 2022).

Theory Meets Practice

In practice, the observed errors are highly aligned with our theoretical bounds. The expected error of the gradient estimator $\widehat{\nabla} f_k^\delta(x)$ can be bounded by

$$\begin{aligned}\mathbb{E}\left[\underbrace{\left\|\widehat{\nabla} f_k^\delta(x) - \nabla f(x)\right\|}_{\text{“error”}}\right] &\leq \mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right]\right\|\right] + \mathbb{E}\left[\left\|\mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x)\right\|\right] \\ &\leq \sqrt{\underbrace{\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right]\right\|^2\right]}_{\text{“variance”}}} + \underbrace{\left\|\mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x)\right\|}_{\text{“bias”}},\end{aligned} \tag{3}$$

which implies that the variance is critical in bridging the theoretical bias bounds and the practical performance. In fact, the variance bound is highly aligned with the empirical error, as illustrated in Figure 1.

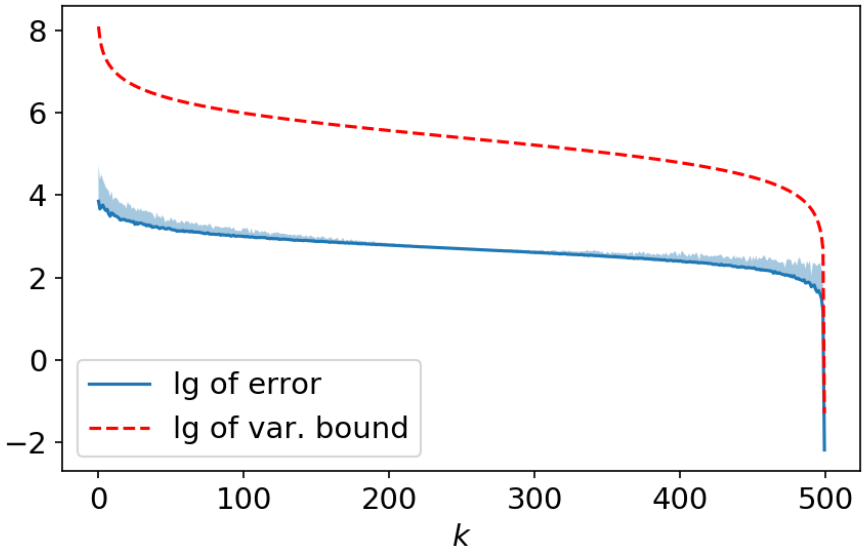


Figure 1: Errors (Eq. 3) of gradient estimators $\widehat{\nabla} f_k^{\delta}(x)$ with k ranging from 1 to n, in base-10 log-scale. Here $\delta = 0.01$ and $n = 500$. The underlying test function is $f(x) = \exp((x_1 - 1)(x_2 + 2)) + \sum_{j=1}^{500} \sin(x_j)$, where x_j denotes the j-th component of vector x. The gradient is estimated at $x = 0$. The solid blue curve plots the errors of gradient estimator, and is averaged over 10 runs. The shaded area above the solid curve shows 10 times standard deviation of the errors in logarithmic scale. The dashed red curve, as a function of k, is $c(k) = \lg\left(\|\nabla f(x)\|^2 \left(\frac{n}{k} - 1\right) + \delta^2 \left(\frac{n^2}{k} - n\right) \|\nabla f(x)\| + \frac{\delta^4 n^2}{k}\right)$, which is the base-10 log of of variance bound for the gradient estimators (up to constants). The shapes of the two curves are highly aligned. More details are in Section 5.

"Better-Than-Definition" Accuracy

In most numerical analysis textbooks, the default finite-difference gradient/Hessian estimator is the entry-wise estimator: We perform a 1-dimensional finite difference estimation for each entry of the gradient/Hessian, and gather all entries to output a gradient/Hessian estimator. Often times this method is considered the "definition" for the task of zeroth order gradient/Hessian estimation.

As one would naturally expect, it is hard, if possible, to outperform this "definition" in an environment where (i) one can sample as many zeroth-order function evaluations as she wants, and (ii) all function evaluations are noise-free. Surprisingly, when $k = n$, our estimators (Eq. 1) and (Eq. 2) *can outperform* the entry-wise estimators (the "definition"). Some numerical comparisons between our estimators and the entry-wise estimators are in Table 1, and more details can be found in Section 5.

Note that this observation is not in conflict with previous works (Flaxman et al., 2005; Wang et al., 2021; Nesterov and Spokoiny, 2017; Balasubramanian and Ghadimi, 2021; Wang, 2022), since they focus on scenarios where either (i) one has to estimate the gradient/Hessian with number of samples much smaller than the dimensionality of the space (Flaxman et al., 2005; Wang et al., 2021), or (ii) there is noise in the zeroth-order function evaluations (Nesterov and Spokoiny, 2017; Balasubramanian and Ghadimi, 2021; Wang, 2022).

Related Works

Zeroth order optimization is a central topic in many fields (e.g., Nelder and Mead, 1965; Goldberg and Holland, 1988; Conn et al., 2009; Nemirovski et al., 2009; Shahriari et al., 2015). Among many zeroth order optimization mechanisms, a classic and prosperous line of works focuses on estimating higher order derivatives using zeroth order information.

Previously, Flaxman et al. (2005) studied the stochastic gradient estimator using a single-point function evaluation for the purpose of bandit learning. Duchi et al. (2015) studied stabilization of the stochastic gradient estimator via two-points (or multi-points) evaluations. Nesterov and Spokoiny (2017); Balasubramanian and Ghadimi (2021) studied gradient/Hessian estimators us-

Table 1: The errors of gradient estimators listed against finite-difference granularity δ. The first row shows δ; The second row shows errors of $\widehat{\nabla} f_n^\delta(x)$ ($n = 500$ is the dimension); The third row shows errors of the entry-wise estimator (More details in Section 5). The error of an estimator is its distance to the true gradient (in Euclidean norm). Errors of $\widehat{\nabla} f_n^\delta(x)$ are in average $\pm$ standard derivation format, where each average and standard deviation gather information from 10 runs. The entry-wise estimator is not random and its error is computed in a single run. The underlying test function is the same as the test function in Figure 1, and $x = 0$ is used for all evaluations.

δ	0.1	0.01	0.001
Stiefel's sampling errors	2.8e-4±4.0e-6	2.8e-6±1.0e-07	2.9e-08±6.4e-10
Entry-wise errors	3.8e-2	3.7e-4	3.7e-6

ing Gaussian smoothing, and investigated downstream optimization methods using the estimated gradient. In particular, Balasubramanian and Ghadimi (2021) studied the zeroth order Hessian estimators via the Stein's identity (Stein, 1981) and applied the estimator to cubic regularized Newton's method (Nesterov and Polyak, 2006). In addition to the above mentioned works, comparison-based gradient estimator has also been considered by Cai et al. (2022), which follows from a rich line of works in information theory (e.g., Raginsky and Rakhlin, 2011; Jamieson et al., 2012; Plan and Vershynin, 2012, 2014).

Perhaps the most relevant works are (Flaxman et al., 2005) for gradient estimators, and (Wang, 2022) for Hessian estimators. As for gradient estimators, (Eq. 1) includes the estimator by Flaxman et al. (2005) as a special case when $k = 1$. We show that variance of the gradient estimator can be significantly reduced as we increase k. Also, we provide an $\mathcal{O}(\delta^2)$ bias bound for the gradient estimator, which improves the $\mathcal{O}(\delta)$ bias bound by (Flaxman et al., 2005). The comparison with (Wang, 2022) is more subtle, since the estimation tasks are conducted in different spaces. While the primary goal of (Wang, 2022) is a curvature-dependent bias bound over Riemannian manifolds, the primary goals of the current paper is variance reduction and refined bias bounds in Euclidean spaces. Similar to the gradient case, we propose variance reduction methods for the Hessian estimators, and provide improved bias bounds for the Hessian estimators, both in Euclidean spaces.

.

2 Preliminaries

We list here some preliminaries, assumptions, and notations before proceeding to subsequent sections. Throughout the paper, we restrict our attention to real-valued functions defined over $\mathbb{R}^n$. The letter n is reserved for the dimension of the space, unless otherwise specified. Also, $\|\cdot\|$ is reserved for the Euclidean norm when applied to vectors, and reserved for the spectral norm when applied to matrices (or symmetric tensors). For a function that is m-times continuously differentiable, let $\partial^m f$ denote the bundle of the total derivatives of f. More specifically, for any $x \in \mathbb{R}^n$, $\partial^m f(x)$ is an m-order symmetric multi-linear form (a tensor). In particular, $\partial^1 f = \nabla f$ and $\partial^2 f = \nabla^2 f$. For this multi-linear form $\partial^m f(x)$, which maps m vectors in $\mathbb{R}^n$ to $\mathbb{R}$, we write $\partial^m f(x)[v]$ as a shorthand for $\partial^m f(x)[\underbrace{v, v, \cdots, v}_{m\text{times}}]$.

For different tasks, we make different assumptions about the smoothness of the function, which is described by the following (k, L)-smoothness terminology.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called (k, L)-smooth ($k \in \mathbb{N}_+$, $L > 0$) if it is k-times continuously differentiable, and*

$$\|\partial^{k-1} f(x) - \partial^{k-1} f(x')\| \le L, \quad \forall x, x' \in \mathbb{R}^n,$$

where $\|F\| := \sup_{v \in \mathbb{R}^n, \|v\|=1} F[v]$ is the spectral norm for the symmetric multi-linear form F.

Throughout, $\|\cdot\|$ is reserved for the spectral norm when applied to symmetric multi-linear forms (including symmetric matrices). Although the smoothness quantification in Definition 1 is global, a local version can be similarly defined, and all subsequent results can be obtained, using similar arguments. All (k, L)-smooth functions satisfy the following proposition.

Proposition 1. *If f is (k, L)-smooth, then*

$$\|\partial^k f(x)\| \leq L, \quad \forall x \in \mathbb{R}^n.$$

Proof. See Appendix. □

We use $\mathbb{S}^{n-1}$ (resp. $\mathbb{B}^n$) to denote the unit sphere (resp. ball) in $\mathbb{R}^n$. Also, given $\delta > 0$, $\delta\mathbb{S}^{n-1}$ (resp. $\delta\mathbb{B}^n$) refers to the origin centered sphere (resp. ball) of radius δ in $\mathbb{R}^n$. Several useful identities are stated below in Propositions 2, Proposition 3, and Proposition 4. References for the following propositions include (Nesterov and Spokoiny, 2017; Wang, 2022; Cai et al., 2022). Their proofs are included in the appendix.

Proposition 2. *Let v be a vector uniformly randomly sampled from $\mathbb{S}^{n-1}$. Then it holds that*

$$\mathbb{E}\left[vv^\top\right] = \frac{1}{n} I,$$

where I is the identity matrix (of size $n \times n$).

Proof. See Appendix. □

Proposition 3. *Let v be a vector uniformly randomly sampled from $\mathbb{S}^{n-1}$, and let v_i be the i-th component of v. Then it holds that*

- $\mathbb{E}\left[v_i^4\right] = \frac{3}{n^2+2n}$ *for all* $i = 1, 2, \cdots, n$;
- $\mathbb{E}\left[v_i^2 v_j^2\right] = \frac{1}{n^2+2n}$ *for all* $i, j = 1, 2, \cdots, n$ *and* $i \neq j$;
- $\mathbb{E}\left[v_i v_j v_k v_l\right] = 0$ *for all* $i, j, k, l = 1, 2, \cdots, n$ *and* $i \notin \{j, k, l\}$.

Proof. See Appendix. □

Proposition 4. *Let v, w be two independent vectors uniformly randomly sampled from $\mathbb{S}^{n-1}$, and let A be a symmetric matrix. It holds that*

$$\mathbb{E}\left[\left(v^\top A w\right) v w^\top\right] = \mathbb{E}\left[\left(v^\top A w\right) w v^\top\right] = \frac{1}{n^2} A.$$

Proof. See Appendix. □

3 Gradient Estimation

Since both the gradient and Hessian estimators (Eq. 1) and (Eq. 2) use random directions sampled from the Stiefel's manifold, we first describe the sampling process for generating such random directions. This sampling procedure is summarized in Algorithm 1.

Algorithm 1 Gram-Schmidt Sampling $\texttt{GS}(n, k)$

1: **Input**: Dimension n. Number of vectors k.
2: Sample a random matrix $U \in \mathbb{R}^{n \times k}$ such that $U_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$.
3: Let $[v_1, v_2, \cdots, v_k] = U\left(U^\top U\right)^{-1/2}$.
   /* With probability 1, $[v_1, v_2, \cdots, v_k]$ is well-defined. */
4: **Output:** $\texttt{GS}(n, k) = v_1, v_2, \cdots, v_k$.

The marginal distribution for any vector from Algorithm 1 is a uniform distribution over $\mathbb{S}^{n-1}$, as summarized in Proposition 2.

Proposition 5 (Chikuse (2003))**.** *Let $v_1, v_2, \cdots, v_k$ be vectors sampled from Algorithm 1. The marginal distribution for any v_i is uniform over $\mathbb{S}^{n-1}$.*

Intuitively, Proposition 5 is due to the fact that the uniform measure over the Stiefel's manifold $\mathrm{St}(n,k)$ (the Hausdorff measure with respect to the Frobenius inner product of proper dimension, which is rotation-invariant) can be decomposed into a wedge product of the spherical measure over $\mathbb{S}^{n-1}$ and the uniform measure over $\mathrm{St}(n-1,k-1)$ (Chikuse, 2003). One can link the decomposition of measure to the following sampling process. We first sample v_1 uniformly from $\mathbb{S}^{n-1}$, then v_2 uniformly from $\mathbb{S}^{n-1}\cap\{v_1\}^\perp$ and so on. Thus Algorithm 1 is named Gram-Schmidt sampling. By symmetry, the marginal distribution for any v_i generated from Algorithm 1 is uniform over $\mathbb{S}^{n-1}$, as stated in Proposition 5. We refer the readers to Chapters 1 & 2 in (Chikuse, 2003) for more details on distribution over Stiefel's manifolds.

Using the vectors sampled from Algorithm 1, we define the gradient and Hessian estimators (Eq. 1) and (Eq. 2). This sampling trick can significantly reduce variance.

3.1 Variance of Gradient Estimator

Define $u_i := \frac{n}{2\delta}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i \approx n v_i v_i^\top \nabla f(0)$, where the approximate equality follows from Taylor's theorem. If $[v_1, v_2, \cdots, v_k]$ is sampled from Algorithm 1, $\{u_1, u_2, \cdots, u_k\}$ are mutually perpendicular and we have

$$
\begin{aligned}
&\frac{1}{k^2}\mathbb{E}\left[\left\|\sum_{i=1}^k u_i\right\|^2\right] - \frac{1}{k^2}\left\|\mathbb{E}\left[\sum_{i=1}^k u_i\right]\right\|^2 \\
=\ &\frac{1}{k^2}\mathbb{E}\left[\sum_{i,j:1\le i,j\le k, i\neq j} u_i^\top u_j\right] + \frac{1}{k^2}\left(\mathbb{E}\left[\sum_{i=1}^k \|u_i\|^2\right] - \left\|\mathbb{E}\left[\sum_{i=1}^k u_i\right]\right\|^2\right) \\
\overset{①}{=}\ &\frac{1}{k^2}\left(\mathbb{E}\left[\sum_{i=1}^k \|u_i\|^2\right] - \left\|\mathbb{E}\left[\sum_{i=1}^k u_i\right]\right\|^2\right) \\
\overset{②}{\approx}\ &\frac{1}{k^2}\left(n^2\mathbb{E}\left[\sum_{i=1}^k \nabla f(0)^\top v_i v_i^\top \nabla f(0)\right] - \left\|n\mathbb{E}\left[\sum_{i=1}^k v_i v_i^\top \nabla f(0)\right]\right\|^2\right) \\
\overset{③}{=}\ &\left(\frac{n}{k}-1\right)\|\nabla f(0)\|^2,
\end{aligned}
$$

where ① uses orthogonality of u_i and u_j for $i\neq j$, ② uses $u_i \approx n v_i v_i^\top \nabla f(0)$, and ③ uses Proposition 2. With a Taylor expansion with higher precision, we have the following theorem.

Theorem 1. *If f is $(3, L_3)$-smooth, the variance of the gradient estimator for f satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla} f_k^\delta(x) - \mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right]\right\|^2\right] \\
\le\ &\left(\frac{n}{k}-1\right)\|\nabla f(x)\|^2 + \frac{L_3\delta^2}{3}\left(\frac{n^2}{k}-n\right)\|\nabla f(x)\| + \frac{L_3^2 n^2\delta^4}{36k}, \quad \forall x\in\mathbb{R}^n.
\end{aligned}
$$

Proof. Without loss of generality, let $x=0$. By Taylor expansion, we know that for any $v_i\in\mathbb{S}^{n-1}$ and small δ,

$$
\frac{1}{2}\big(f(\delta v_i) - f(-\delta v_i)\big) = \delta v_i^\top \nabla f(0) + \frac{\delta^3\left(\partial^3 f(z_i)[v_i] + \partial^3 f(z_i')[v_i]\right)}{12},
$$

where z_i, z_i' depend on v_i and δ. For simplicity, let $R_i = \frac{\delta^3\left(\partial^3 f(z_i)[v_i] + \partial^3 f(z_i')[v_i]\right)}{12}$, and $|R_i| \le \frac{L_3\delta^3}{6}$ for all $i = 1, 2, \cdots, k$.

For any i, k, n, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right\|^2\right]-\left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]\right\|^2\\
&=\mathbb{E}\left[\left\|\left(\delta v_i^\top\nabla f(0)+R_i\right)v_i\right\|^2\right]-k\left\|\mathbb{E}\left[\left(\delta v_i^\top\nabla f(0)+R_i\right)v_i\right]\right\|^2\\
&\overset{①}{=}\mathbb{E}\left[\delta^2\nabla f(0)^\top v_i v_i^\top\nabla f(0)+2\delta\nabla f(0)^\top v_i R_i+R_i^2\right]-k\left\|\mathbb{E}\left[\delta v_i v_i^\top\nabla f(0)+R_i v_i\right]\right\|^2.
\end{aligned}$$

Since $\mathbb{E}\left[v_i v_i^\top\right]=\frac{1}{n}I$, ① gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right\|^2\right]-\left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]\right\|^2\\
&=\frac{\delta^2}{n}\|\nabla f(0)\|^2+2\delta\nabla f(0)^\top\mathbb{E}\left[R_i v_i\right]+\mathbb{E}\left[R_i^2\right]-k\left\|\mathbb{E}\left[\frac{\delta}{n}\nabla f(0)+R_i v_i\right]\right\|^2\\
&=\left(\frac{\delta^2}{n}-\frac{\delta^2 k}{n^2}\right)\|\nabla f(0)\|^2+\left(2\delta-\frac{2\delta k}{n}\right)\nabla f(0)^\top\mathbb{E}\left[R_i v_i\right]+\left(\mathbb{E}\left[R_i^2\right]-k\mathbb{E}\left[R_i v_i\right]^\top\mathbb{E}\left[R_i v_i\right]\right)\\
&\overset{②}{\leq}\left(\frac{\delta^2}{n}-\frac{\delta^2 k}{n^2}\right)\|\nabla f(0)\|^2+\frac{L_3\delta^4}{3}\left(1-\frac{k}{n}\right)\|\nabla f(0)\|+\frac{L_3^2\delta^6}{36}.
\end{aligned}$$

For the variance of the gradient estimator, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla}f_k^\delta(0)-\mathbb{E}\left[\widehat{\nabla}f_k^\delta(0)\right]\right\|^2\right]\\
&=\mathbb{E}\left[\left\|\widehat{\nabla}f_k^\delta(0)\right\|^2\right]-\left\|\mathbb{E}\left[\widehat{\nabla}f_k^\delta(0)\right]\right\|^2\\
&=\mathbb{E}\left[\left\|\frac{n}{2\delta k}\sum_{i=1}^k\left(f(\delta v_i)-f(\delta v_i)\right)v_i\right\|^2\right]-\left\|\frac{n}{2\delta k}\sum_{i=1}^k\mathbb{E}\left[\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]\right\|^2\\
&\overset{③}{=}\frac{n^2}{\delta^2k^2}\sum_{i=1}^k\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i)-f(\delta v_i)\right)v_i\right\|^2\right]\\
&\quad-\frac{n^2}{4\delta^2k^2}\sum_{i,j=1}^k\mathbb{E}\left[\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]^\top\mathbb{E}\left[\left(f(\delta v_j)-f(-\delta v_j)\right)v_j\right],
\end{aligned}$$

where the last equation follows from the orthonormality of $\{v_1, v_2.\cdots, v_k\}$. By Proposition 5, we know that $\mathbb{E}\left[\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]=\mathbb{E}\left[\left(f(\delta v_j)-f(-\delta v_j)\right)v_j\right]$ for all $i, j=1, 2, \cdots, k$. Thus ③ gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla}f_k^\delta(0)-\mathbb{E}\left[\widehat{\nabla}f_k^\delta(0)\right]\right\|^2\right]\\
&\overset{④}{=}\frac{n^2}{\delta^2k^2}\sum_{i=1}^k\left(\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right\|^2\right]-\left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]\right\|^2\right)
\end{aligned}$$

Combining ② and ④ gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\nabla}f_k^\delta(0)-\mathbb{E}\left[\widehat{\nabla}f_k^\delta(0)\right]\right\|^2\right]\\
&=\frac{n^2}{\delta^2k^2}\sum_{i=1}^k\left(\mathbb{E}\left[\left\|\frac{1}{2}\left(f(\delta v_i)-f(\delta v_i)\right)v_i\right\|^2\right]-\left\|\mathbb{E}\left[\frac{\sqrt{k}}{2}\left(f(\delta v_i)-f(-\delta v_i)\right)v_i\right]\right\|^2\right)\\
&\leq\left(\frac{n}{k}-1\right)\|\nabla f(0)\|^2+\frac{L_3\delta^2}{3}\left(\frac{n^2}{k}-n\right)\|\nabla f(0)\|+\frac{L_3^2n^2\delta^4}{36k}.
\end{aligned}$$

□

3.2 Bias of Gradient Estimator

While the estimator $\widehat{\nabla} f_k^\delta$ can significantly reduce variance, it does not sacrifice any bias accuracy. There has been a sequence of works on bias of gradient estimators of (Eq. 1) or alternatives of (Eq. 1) (Flaxman et al., 2005; Nesterov and Spokoiny, 2017; Wang et al., 2021). In this section, we provide a refined analysis on the bias, and present a bias bound of order $\mathcal{O}\left(\min\{\delta, \delta^2\}\right)$.

Theorem 2. *The gradient estimator $\widehat{\nabla} f_k^\delta$ satisfies*

(a) If f is $(1, L_1)$-smooth, then for all $x \in \mathbb{R}^n$, $\left\|\mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x)\right\| \leq \frac{L_1 n \delta}{n+1}$.

(b) If f is $(4, L_4)$-smooth, then for all $x \in \mathbb{R}^n$, the bias of gradient estimator satisfies

$$\left\|\mathbb{E}\left[\widehat{\nabla} f_k^\delta(x)\right] - \nabla f(x)\right\| \leq \frac{\delta^2}{2n}\sqrt{\sum_{i=1}^n \left(\sum_{j=1}^n F_{jji}\right)^2} + \frac{\delta^3 L_4 n}{24},$$

where F_{ijl} denotes the (i, j, l)-component of $\partial^3 f(x)$.

Item (a) in Theorem 2 is due to Flaxman et al. (2005). This fact can be viewed as a consequence of the Stokes' theorem, or the fundamental theorem of (geometric) calculus, or the divergence theorem. A proof for item (a) using divergence theorem is in Appendix for completeness.

Item (b) is a refined analysis of Taylor expansion and spherical random projection properties. More specifically, we expand $f(x + \delta v_i)v_i$ up to fourth order and repeatedly exploit Propositions 2 and 3. Below we provide a detailed proof for item (b).

Proof of Theorem 2(b). Without loss of generality, let $x = 0$. By Proposition 5, we know that $\mathbb{E}[\widehat{\nabla} f_k^\delta(x)] = \frac{n}{\delta}\mathbb{E}[f(\delta v)v]$, where v is uniformly sampled from $\mathbb{S}^{n-1}$. Taylor expansion gives that

$$f(\delta v) = f(0) + \nabla f(0)^\mathsf{T}(\delta v) + \frac{1}{2}(\delta v)^\mathsf{T}\partial^2 f(0)[\delta v] + \frac{1}{6}\partial^3 f(0)[\delta v] + \frac{1}{24}\partial^4 f(\xi(\delta v))[\delta v],$$

where we use notation $\xi(\delta v)$ to indicate that ξ is a function of δv.

Therefore, the expectation can be written as

$$\begin{aligned}\mathbb{E}[f(\delta v)v] \overset{①}{=} &\mathbb{E}\left[f(0)v\right] + \delta\mathbb{E}\left[vv^\mathsf{T}\nabla f(0)\right] + \frac{\delta^2}{2}\mathbb{E}\left[\left(v^\mathsf{T}\nabla^2 f(0)v\right)v\right] \\ &+ \frac{\delta^3}{6}\mathbb{E}\left[\partial^3 f(0)[v]v\right] + \frac{\delta^4}{24}\mathbb{E}\left[\partial^4 f(\xi(\delta v))[v]v\right].\end{aligned}$$

The first and the third term in ① are zero, since the expectation is with respect to a uniform distribution over the unit sphere and the integrands are odd. Thus we have

$$\mathbb{E}\left[f(0)v\right] = \mathbb{E}\left[\left(v^\mathsf{T}\nabla^2 f(0)v\right)v\right] \overset{②}{=} 0$$

For the second term in ①, Proposition 5 gives

$$\mathbb{E}\left[vv^\mathsf{T}\nabla f(0)\right] \overset{③}{=} \frac{1}{n}\nabla f(0).$$

For the forth term, we denote $\partial^3(f(0))$ as F for simplicity. Then the i-th component of $\partial^3 f(0)[v]v$ is $\sum_{j,l,p} F_{jlp}v_j v_l v_p v_i$. By Proposition 3, we have

$$\begin{aligned}\left(\mathbb{E}\left[\partial^3 f(0)[v]v\right]\right)_i &= \sum_{j,l,p} F_{jlp}\mathbb{E}\left[v_j v_l v_p v_i\right] \\ &= \sum_{j\neq i}\frac{1}{n^2+2n}\left(F_{jji} + F_{jij} + F_{ijj}\right) + \frac{3}{n^2+2n}F_{iii} \\ &= \frac{3}{n^2+2n}\sum_{j=1}^n F_{jji}.\end{aligned}$$

Thus it holds that

$$\frac{\delta^3}{6}\mathbb{E}\left[\partial^3 f(0)[v]v\right] \overset{④}{=} \frac{\delta^3}{2n^2+4n}\sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} F_{jji}\right)^2}.$$

For the fifth term, Proposition 1 gives that $\|\partial^4 f(x)\| \leq L_4$, so we have

$$\left\|\frac{\delta^4}{24}\mathbb{E}\left[\partial^4 f(\xi(\delta v))[\delta v]v\right]\right\| \overset{⑤}{\leq} \frac{\delta^4 L_4}{24}.$$

Now combining ② ③ ④ ⑤, we arrive at the upper bound

$$\left\|\mathbb{E}[f(\delta v)v] - \frac{\delta}{n}\nabla f(0)\right\| \leq \frac{\delta^3}{2n^2+4n}\sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} F_{jji}\right)^2} + \frac{\delta^4 L_4}{24}$$

and

$$\left\|\frac{n}{\delta}\mathbb{E}[f(\delta v)v] - \nabla f(0)\right\| \leq \frac{\delta^2}{2n}\sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{n} F_{jji}\right)^2} + \frac{\delta^3 L_4 n}{24}.$$

□

4 Hessian Estimation

Previously, Wang (2022) introduced the Hessian estimator (Eq. 2) with $k = 1$, over Riemannian manifolds. Similar to the gradient case, when we sample orthogonal frames for estimation, the variance can be reduced. As previously discussed, the variance of $\widehat{\mathrm{H}} f_k^\delta(x)$ for a smooth f defined in $\mathbb{R}^n$ is of order

$$\mathcal{O}\left(\left(\frac{n^2}{k^2} - 1\right) + \left(\frac{n^4}{k^2} - k^2\right)\delta^2 + \frac{n^4\delta^4}{k^2}\right).$$

Similar to its gradient estimator counterpart, variance of $\widehat{\mathrm{H}} f_k^\delta(x)$ eventually goes to $\mathcal{O}\left(\delta^4\right)$ when $k = n$. However, the task of Hessian estimation is harder, since:

1. When k is small compare to n, the variance is of order $\mathcal{O}\left(n^2\right)$, which is worse than that for the gradient estimator.
2. The estimator $\widehat{\mathrm{H}} f_k^\delta(x)$ requires $\Theta(k^2)$ samples, which means it takes $\Theta(n^2)$ samples to reach a negligible variance.

4.1 Variance of Hessian Estimator

Similar to that for gradient estimators, the variance of Hessian estimator is bounded via high-order Taylor expansion and random projection arguments. A high-precision bound on the variance is in Theorem 3.

Theorem 3. *If the underlying function is $(4, L_4)$-smooth and $(6, L_6)$-smooth, then the Hessian estimator $\widehat{\mathrm{H}} f_k^\delta(x)$ (Eq. 2) with $\delta > 0$ satisfies, for all $x \in \mathbb{R}^n$*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\widehat{\mathrm{H}} f_k^\delta(x) - \mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(x)\right]\right\|_F^2\right] \\
\leq\ &\left\|\nabla^2 f(x)\right\|_F^2\left(\frac{n^2}{k^2} - 1\right) + 2\delta^2 L_4\left\|\nabla^2 f(x)\right\|\left(\frac{n^4}{k^2} - n^2\right) + \mathcal{O}\left(\left(L_6 n^2\|\nabla^2 f(x)\| + \frac{n^4 L_4^2}{k^2}\right)\delta^4\right).
\end{aligned}$$

Proof. Without loss of generality, let $x = 0$. Using Taylor series expansion, for any v_i, w_j, we have

$$\begin{aligned}&\frac{1}{4}\left(f(\delta v_i + \delta w_j) - f(\delta v_i - \delta w_j) - f(-\delta v_i + \delta w_j) + f(-\delta v_i - \delta w_j)\right)\\ &= \delta^2 v_i^\top \nabla^2 f(0) w_j + \delta^4 R_{ij} + \delta^6 S_{ij},\end{aligned}$$

where

$$\begin{aligned}R_{ij} &= \frac{1}{96}\left(\partial^4 f(0)[v_i + w_j] - \partial^4 f(0)[-v_i + w_j] - \partial^4 f(0)[v_i - w_j] + \partial^4 f(0)[-v_i - w_j]\right),\\ S_{ij} &= \frac{1}{2880}\left(\partial^6 f(z_1)[v_i + w_j] - \partial^6 f(z_2)[-v_i + w_j] - \partial^6 f(z_3)[v_i - w_j] + \partial^6 f(z_4)[-v_i - w_j]\right)\end{aligned}$$

with z_1, z_2, z_3, z_4 depending on v_i, w_j and δ. Since f is $(4, L_4)$-smooth and $(6, L_6)$-smooth, $|R_{ij}| \le \frac{2L_4}{3}$ and $|S_{ij}| \le \frac{4L_6}{45}$.

For the Frobenius norm of the Hessian estimator, we have

$$\begin{aligned}\left\|\widehat{\mathrm{H}} f_k^\delta(0)\right\|_F^2 &= \frac{n^4}{4k^4\delta^4}\left\|\sum_{i,j=1}^k \left(\frac{\delta^2}{2} v_i^\top \nabla^2 f(0) w_j + \frac{\delta^2}{2} w_j^\top \nabla^2 f(0) v_i + \delta^4 R_{ij} + \delta^6 S_{ij}\right)(v_i w_j^\top + w_j v_i^\top)\right\|_F^2\\ &\le \frac{n^4}{2k^4}\left\|\sum_{i,j=1}^k \left(\frac{1}{2} v_i^\top \nabla^2 f(0) w_j + \frac{1}{2} w_j^\top \nabla^2 f(0) v_i + \delta^2 R_{ij} + \delta^4 S_{ij}\right) v_i w_j^\top\right\|_F^2\\ &\quad + \frac{n^4}{2k^4}\left\|\sum_{i,j=1}^k \left(\frac{1}{2} v_i^\top \nabla^2 f(0) w_j + \frac{1}{2} w_j^\top \nabla^2 f(0) v_i + \delta^2 R_{ij} + \delta^4 S_{ij}\right) w_j v_i^\top\right\|_F^2\\ &\overset{①}{\le} \frac{n^4}{4k^4}\left\|\sum_{i,j=1}^k \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij}\right) v_i w_j^\top\right\|_F^2\\ &\quad + \frac{n^4}{4k^4}\left\|\sum_{i,j=1}^k \left(w_j^\top \nabla^2 f(0) v_i + \delta^2 R_{ij} + \delta^4 S_{ij}\right) v_i w_j^\top\right\|_F^2\\ &\quad + \frac{n^4}{4k^4}\left\|\sum_{i,j=1}^k \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij}\right) w_j v_i^\top\right\|_F^2\\ &\quad + \frac{n^4}{4k^4}\left\|\sum_{i,j=1}^k \left(w_j^\top \nabla^2 f(0) v_i + \delta^2 R_{ij} + \delta^4 S_{ij}\right) w_j v_i^\top\right\|_F^2,\end{aligned}$$

where both inequalities use $\|A+B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$ for any matrices A, B of same size. Next, we will bound the expectation of the first term in ①, all other terms can be bounded using similar arguments.

For the first term in ①, we have

$$
\begin{aligned}
&\left\| \sum_{i,j=1}^{k} \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij} \right) v_i w_j^\top \right\|_F^2 \\
&= tr\left(\left(\sum_{i,j=1}^{k} \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij} \right) v_i w_j^\top \right)^\top \left(\sum_{a,b=1}^{k} \left(v_a^\top \nabla^2 f(0) w_b + \delta^2 R_{ab} + \delta^4 S_{ab} \right) v_a w_b^\top \right) \right) \\
&= \sum_{i,j=1}^{k} \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij} \right)^2 \\
&\overset{②}{=} \sum_{i,j=1}^{k} \left(v_i^\top \nabla^2 f(0) w_j \right)^2 + \sum_{i,j=1}^{k} 2 \left(v_i^\top \nabla^2 f(0) w_j \right) \left(\delta^2 R_{ij} + \delta^4 S_{ij} \right) + \sum_{i,j=1}^{k} \left(\delta^2 R_{ij} + \delta^4 S_{ij} \right)^2,
\end{aligned}
$$

where the second last line uses orthogonality of v_i's and w_j's.

Since $\mathbb{E}\left[\left(v_i^\top \nabla^2 f(0) w_j \right)^2 \right] = \mathbb{E}\left[tr\left(v_i v_i^\top \nabla^2 f(0) w_j w_j^\top \nabla^2 f(0) \right) \right] = \frac{1}{n^2} \|\nabla^2 f(0)\|_F^2$, taking expectation on both sides of ② gives,

$$
\begin{aligned}
&\mathbb{E}\left[\left\| \sum_{i,j=1}^{k} \left(v_i^\top \nabla^2 f(0) w_j + \delta^2 R_{ij} + \delta^4 S_{ij} \right) v_i w_j^\top \right\|_F^2 \right] \\
&\leq \frac{k^2}{n^2} \left\| \nabla^2 f(0) \right\|_F^2 + 2\delta^2 \sum_{i,j=1}^{k} \mathbb{E}\left[v_i^\top \nabla^2 f(0) w_j R_{ij} \right] + 2k^2 \left\| \nabla^2 f(0) \right\| \frac{4L_6 \delta^4}{45} + k^2 \left(\frac{2L_4 \delta^2}{3} + \frac{4L_6 \delta^4}{45} \right)^2.
\end{aligned}
$$

Also, we have

$$
\begin{aligned}
&\left\| \mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(0) \right] \right\|_F^2 \\
&= \frac{n^4}{4k^4} \left\| \mathbb{E}\left[\sum_{i,j=1}^{k} \left(\frac{1}{2} v_i^\top \nabla^2 f(0) w_j + \frac{1}{2} w_j^\top \nabla^2 f(0) v_i + \delta^2 R_{ij} + \delta^4 S_{ij} \right) \left(v_i w_j^\top + w_j v_i^\top \right) \right] \right\|_F^2 \\
&\overset{③}{=} \frac{n^4}{4k^4} \left\| \frac{2k^2}{n^2} \nabla^2 f(0) + \mathbb{E}\left[\sum_{i,j=1}^{k} \left(\delta^2 R_{ij} + \delta^4 S_{ij} \right) \left(v_i w_j^\top + w_j v_i^\top \right) \right] \right\|_F^2 \\
&\geq \left\| \nabla^2 f(0) \right\|_F^2 + 2 \frac{n^4}{4k^4} \cdot \frac{2k^2}{n^2} tr\left(\nabla^2 f(0) \mathbb{E}\left[\sum_{i,j=1}^{k} \left(\delta^2 R_{ij} + \delta^4 S_{ij} \right) \left(v_i w_j^\top + w_j v_i^\top \right) \right] \right) \\
&\geq \left\| \nabla^2 f(0) \right\|_F^2 + \frac{2n^2 \delta^2}{k^2} \sum_{i,j=1}^{k} \mathbb{E}\left[v_i^\top \nabla^2 f(0) w_j R_{ij} \right] - 2n^2 \left\| \nabla^2 f(0) \right\| \frac{4L_6 \delta^4}{45},
\end{aligned}
$$

where ③ uses Proposition 4.

Collecting terms gives

$$
\begin{aligned}
&\left\|\widehat{\mathrm{H}} f_k^\delta(0)\right\|_F^2 - \left\|\mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(0)\right]\right\|_F^2 \\
\leq\ & \frac{n^4}{k^4}\left(\frac{k^2}{n^2}\left\|\nabla^2 f(0)\right\|_F^2 + 2\delta^2 \sum_{i,j=1}^k \mathbb{E}\left[v_i^\top \nabla^2 f(0) w_j R_{ij}\right] + 2k^2\left\|\nabla^2 f(0)\right\| \frac{4L_6\delta^4}{45} + k^2\left(\frac{2L_4\delta^2}{3} + \frac{4L_6\delta^4}{45}\right)^2\right) \\
&- \left\|\nabla^2 f(0)\right\|_F^2 - \frac{2n^2\delta^2}{k^2}\sum_{i,j=1}^k \mathbb{E}\left[v_i^\top \nabla^2 f(0) w_j R_{ij}\right] + 2n^2\left\|\nabla^2 f(0)\right\| \frac{4L_6\delta^4}{45} \\
\leq\ & \left\|\nabla^2 f(0)\right\|_F^2\left(\frac{n^2}{k^2} - 1\right) + 2\delta^2 \sum_{i,j=1}^k \mathbb{E}\left[v_i^\top \nabla^2 f(0) w_j R_{ij}\right]\left(\frac{n^4}{k^4} - \frac{n^2}{k^2}\right) + \mathcal{O}\left(\left(L_6 n^2 \|\nabla^2 f(0)\| + \frac{n^4 L_4^2}{k^2}\right)\delta^4\right) \\
\leq\ & \left\|\nabla^2 f(0)\right\|_F^2\left(\frac{n^2}{k^2} - 1\right) + 2\delta^2 L_4\left\|\nabla^2 f(0)\right\|\left(\frac{n^4}{k^2} - n^2\right) + \mathcal{O}\left(\left(L_6 n^2 \|\nabla^2 f(0)\| + \frac{n^4 L_4^2}{k^2}\right)\delta^4\right).
\end{aligned}
$$

□

4.2 Bias of Hessian Estimator

Similar to the gradient case, the Hessian estimator $\widehat{\mathrm{H}} f_k^\delta(x)$ does not sacrifice any bias accuracy. Previously, Wang (2022) showed that the bias of $\widehat{\mathrm{H}} f_k^\delta(x)$ is of order $\mathcal{O}(\delta)$. In particular, Wang (2022) derived a formula for how the local geometry of the Riemannian manifold would affect the bias of the Hessian estimator. In this paper, we focus on providing refined bias bounds in the Euclidean case, and provide an $\mathcal{O}\left(\delta^2\right)$ bias bound, which is stated below in Theorem 4.

Theorem 4. *The Hessian estimator $\widehat{\mathrm{H}} f_k^\delta$ satisfies*

(a) If f is $(2, L_2)$-smooth, then for all $x \in \mathbb{R}^n$, $\left\|\mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(x)\right] - \nabla^2 f(x)\right\| \leq \frac{2nL_2\delta}{n+1}$.

(b) If f is $(5, L_5)$-smooth, then for all $x \in \mathbb{R}^n$, the bias of $\widehat{\mathrm{H}} f_k^\delta$ satisfies

$$\left\|\mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(x)\right] - \nabla^2 f(x)\right\| \leq \frac{\delta^2}{n+2}\left\|\widetilde{F}\right\| + \frac{4\delta^3 L_5 n^2}{15},$$

where $\widetilde{F}$ is an $n \times n$ matrix and $\widetilde{F}_{ij} = \sum_{m=1}^n \left[\partial^4 f(x)\right]_{mmij}$.

Item (a) in Theorem 4 is due to Wang (2022), and its proof via divergence theorem is provided in the Appendix for completeness. Below we provide a proof for item (b).

Proof of Theorem 4(b). Without loss of generality, let $x = 0$. From the definition

$$
\begin{aligned}
\widehat{\mathrm{H}} f_k^\delta(0) = \frac{n^2}{8\delta^2 k^2}\sum_{i,j=1}^k \big(&f(\delta v_i + \delta w_j) - f(-\delta v_i + \delta w_j) - f(\delta v_i - \delta w_j) \\
&+ f(-\delta v_i - \delta w_j)\big) \cdot (v_i w_j^\mathsf{T} + w_j v_i^\mathsf{T}).
\end{aligned}
$$

Note that, by linearity of expectation and symmetry, $\mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(0)\right] = \frac{n^2}{\delta^2}\mathbb{E}\left[f(\delta v + \delta w)(vw^\mathsf{T})\right]$, where v and w are independent and uniformly sampled from $\mathbb{S}^{n-1}$. Taylor expansion gives that

$$
\begin{aligned}
f(\delta v + \delta w) =& f(0) + \partial^1 f(0)[\delta v + \delta w] + \frac{1}{2}\partial^2 f(0)[\delta v + \delta w] + \frac{1}{6}\partial^3 f(0)[\delta v + \delta w] \\
&+ \frac{1}{24}\partial^4 f(0)[\delta v + \delta w] + \frac{1}{120}\partial^5 f(\xi(\delta v + \delta w))[\delta v + \delta w],
\end{aligned}
$$

where we use notation $\xi(\delta v + \delta w)$ to show that ξ is a function of $\delta v + \delta w$.

Therefore, the expectation can be written as

$$\begin{aligned}
&\mathbb{E}\left[f(\delta v+\delta w)(vw^{\mathsf{T}})\right]\\
\overset{①}{=}&\mathbb{E}\left[f(0)(vw^{\mathsf{T}})\right]+\mathbb{E}\left[\partial^1 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]+\frac{1}{2}\mathbb{E}\left[\partial^2 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]\\
&+\frac{1}{6}\mathbb{E}\left[\partial^3 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]+\frac{1}{24}\mathbb{E}\left[\partial^4 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]\\
&+\frac{1}{120}\mathbb{E}\left[\partial^5 f(\xi(\delta v+\delta w))[\delta v+\delta w](vw^{\mathsf{T}})\right]
\end{aligned}$$

Since $\mathbb{E}\left[vw^{\mathsf{T}}\right]\overset{②}{=}0$, the first term in ① equals to 0.

For the second term in ①, the (i,j)-component of $\partial^1 f(0)[v+w]vw^{\mathsf{T}}$ is $\sum_l \partial^1 f(0)_l(v_l+w_l)v_iw_j$. For any l, $v_lv_iw_j$ is odd in w_j. Thus we have

$$\mathbb{E}\left[v_lv_iw_j|v_l=a,v_i=b\right]=0$$

for any a,b, which implies $\mathbb{E}\left[v_lv_iw_j\right]=0$. Similarly $\mathbb{E}\left[w_lv_iv_j\right]=0$. This implies that the second term in ① is zero:

$$\mathbb{E}\left[\partial^1 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]\overset{③}{=}0.$$

Similar arguments show that the forth term equals to 0:

$$\mathbb{E}\left[\partial^3 f(0)[\delta v+\delta w](vw^{\mathsf{T}})\right]\overset{④}{=}0.$$

For the third term, the (i,j)-component of $\partial^1 f(0)[v+w]vw^{\mathsf{T}}$ is $\sum_{p,q}\partial^2 f(0)_{pq}(v_p+w_p)(v_q+w_q)v_iw_j$. It holds that

$$\begin{aligned}
&\mathbb{E}\left[\partial^2 f(0)[v+w](vw^{\mathsf{T}})\right]\\
=&\mathbb{E}\left[\binom{2}{0}\partial^2 f(0)[v](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{2}{1}\partial^2 f(0)[v,w](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{2}{2}\partial^2 f(0)[w,w](vw^{\mathsf{T}})\right]\\
=&\mathbb{E}\left[\binom{2}{1}\partial^2 f(0)[v,w](vw^{\mathsf{T}})\right]\\
=&\frac{2\delta^2}{n^2}\partial^2 f(0),
\end{aligned}$$

where the second last equation follows from that expectation of terms of odd power of v or w is zero, and the last equation uses Proposition 4.

Let $F=\partial^4 f(0)$ for simplicity. For the fifth term, it holds that

$$\begin{aligned}
\mathbb{E}\left[F[v+w](vw^{\mathsf{T}})\right]=&\mathbb{E}\left[\binom{4}{0}F[v](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{4}{1}F[v,v,v,w](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{4}{2}F[v,v,w,w](vw^{\mathsf{T}})\right]\\
&+\mathbb{E}\left[\binom{4}{3}F[v,w,w,w](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{4}{4}F[w](vw^{\mathsf{T}})\right]\\
=&\mathbb{E}\left[\binom{4}{1}F[v,v,v,w](vw^{\mathsf{T}})\right]+\mathbb{E}\left[\binom{4}{3}F[v,w,w,w](vw^{\mathsf{T}})\right]\\
\overset{⑤}{=}&8\mathbb{E}\left[F[v,v,v,w](vw^{\mathsf{T}})\right],
\end{aligned}$$

where the second equation uses the symmetric property to conclude that terms of odd powers are zero (similar to the previous arguments), and the last equation uses symmetry of F and equivalence of v and w.

By Proposition 3, we have that, for any i, j,

$$\begin{aligned}\mathbb{E}\left[\left(F[v,v,v,w](vw^\mathsf{T})\right)_{ij}\right] &= \mathbb{E}\left[\sum_{pqrs} F_{pqrs} v_p v_q v_r w_s v_i w_j\right] \\ &= \mathbb{E}\left[F_{iiij} v_i^4 w_j^2\right] + 3\mathbb{E}\left[\sum_{m:1\le m\le n, m\neq i} F_{mmij} v_i^2 v_m^2 w_j^2\right] \\ &\overset{⑥}{=} \frac{3}{n^2(n+2)} \sum_{m=1}^{n} F_{mmij}.\end{aligned}$$

Combining ⑤ and ⑥ gives

$$\frac{1}{24}\mathbb{E}\left[\left[\partial^4 f(0)[\delta v + \delta w](vw^\mathsf{T})\right]_{ij}\right] \overset{⑦}{\le} \frac{\delta^4}{n^2(n+2)} \sum_{m=1}^{n} F_{mmij},$$

where $F = \partial^4 f(0)$.

For the sixth term in ①, Proposition 1 gives that $\|\partial^5 f(x)\| \le L_5$, so $|\partial^5 f(\xi(\delta v + \delta w))[v+w]| \le 32L_5$. Besides, since $v, w \in \mathbb{S}^{n-1}$, $\|vw^\mathsf{T}\| \le 1$. Therefore, we have

$$\frac{\delta^5}{120}\left\|\mathbb{E}\left[\partial^5 f(\xi(\delta v + \delta w))[v+w] vw^\mathsf{T}\right]\right\| \overset{⑧}{\le} \frac{4\delta^5 L_5}{15}, \tag{4}$$

where $\widetilde{F}$ is Collecting terms from ① ② ③ ④ ⑦ ⑧, we have

$$\left\|\frac{n^2}{\delta^2}\mathbb{E}[f(\delta v + \delta w)(vw^\mathsf{T})] - \nabla^2 f(0)\right\| \le \frac{\delta^2}{n+2}\left\|\widetilde{F}\right\| + \frac{4\delta^3 L_5 n^2}{15},$$

where $\widetilde{F}$ is an $n \times n$ matrix and $\widetilde{F}_{ij} = \sum_{m=1}^{n}\left[\partial^4 f(0)\right]_{mmij}$. □

5 Empirical Studies for Gradient Estimators

In this section, we empirically study the newly introduced gradient estimator. The experiments are divided into three subsections. The first two subsections compare our method with existing methods, and the third subsection empirically verify the theoretical variance bound. To avoid clutter, only some results are listed in the main text, while additional results can be found in the Appendix.

5.1 Comparison with Stochastic Estimators

For the same number of function evaluations specified by k, and finite difference granularity δ, we compare the estimators $\widehat{\nabla} f_k^\delta(x)$ with:

- The estimator via spherical sampling (Flaxman et al., 2005; Wang et al., 2021):

$$\widehat{\nabla} f_{k,S}^\delta(x) := \frac{n}{2k\delta}\sum_{i=1}^{k}\left(f(x+\delta v_i) - f(x-\delta v_i)\right) v_i, \tag{5}$$

  where $v_1, v_2, \cdots, v_k$ are uniformly *i.i.d.* sampled from $\mathbb{S}^{n-1}$.

- The estimator via Gaussian sampling (Nesterov and Spokoiny, 2017):

$$\widehat{\nabla} f_{k,G}^\delta(x) := \frac{\sqrt{n}}{2k\delta}\sum_{i=1}^{k}\left(f\left(x + \frac{\delta v_i}{\sqrt{n}}\right) - f\left(x - \frac{\delta v_i}{\sqrt{n}}\right)\right) v_i, \tag{6}$$

  where $v_1, v_2, \cdots, v_k \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$. Note that the random vectors v_i are divided by $\sqrt{n}$ so that in expectation the step size (granularity) is $\Theta(\delta)$.

Algorithm 2 Comparison-based gradient estimator (Cai et al., 2022)

1: **Input:** Number of random vectors k; Finite difference granularity: δ; Location for evaluation $x \in \mathbb{R}^n$; Comparison oracle for the target function f: $\mathcal{C}_f(x, y) = \text{sign}(f(x) - f(y))$ for all $x, y \in \mathbb{R}^d$; Sparsity parameter: s.
2: Uniformly sample *i.i.d.* vectors $v_1, v_2, \cdots, v_k$ from $\mathbb{S}^{n-1}$.
3: Let $z_i = \mathcal{C}_f(f(x + \delta v_i), f(x))$ for $i = 1, 2, \cdots, k$.
4: **Output:** $\widehat{\nabla} f^{\delta}_{k,C}(x) := \arg\max_{g:\|g\|_1 \le \sqrt{s}, \|g\| \le 1} \sum_{i=1}^{k} z_i v_i^\top g$.

- Comparison-based gradient estimator $\widehat{\nabla} f^{\delta}_{k,C}(x)$, which estimates the normalized gradient. This estimator is defined in Algorithm 2(Cai et al., 2022).

Note that as per its definition, the comparison-based estimator $\widehat{\nabla} f^{\delta}_{k,C}(x)$ does not provide an estimate for $\nabla f(x)$. Instead, it estimates $\frac{\nabla f(x)}{\|\nabla f(x)\|}$. For this reason, the comparison with $\widehat{\nabla} f^{\delta}_{k,C}(x)$, and the comparison with $\widehat{\nabla} f^{\delta}_{k,S}(x)$ and $\widehat{\nabla} f^{\delta}_{k,G}(x)$ are measured under different scales.

All methods are tested using the following function

$$f(x) := \exp((x_1 - 1)(x_2 + 2)) + \sum_{j=1}^{n} \sin(x_j), \tag{7}$$

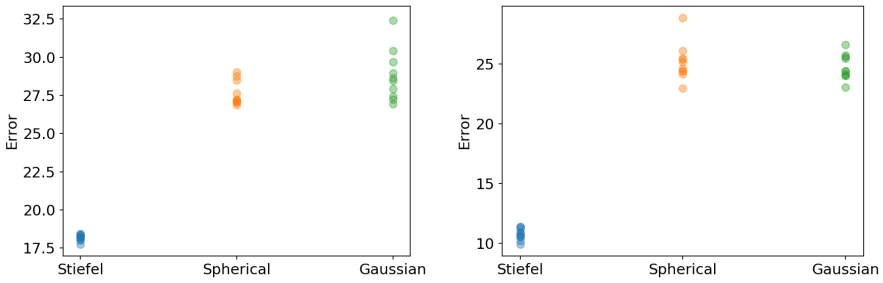
where x_j denotes the j-th component of vector x. Note that, unlikely experiments in some previous works (e.g., Wang, 2022), all function evaluations are *noise-free*. On this test function (Eq. 7), the methods are tested with different choices of k, δ, x. Example comparison between (Eq. 1) and (Eq. 5), (Eq. 6) can be found in Figure 2. Example comparison between (Eq. 1) and $\widehat{\nabla} f^{\delta}_{k,C}(x)$ (Algorithm 2) can be found in Figure 3. More results can be found in the Appendix.

5.2 Comparison with the Entry-wise Estimator

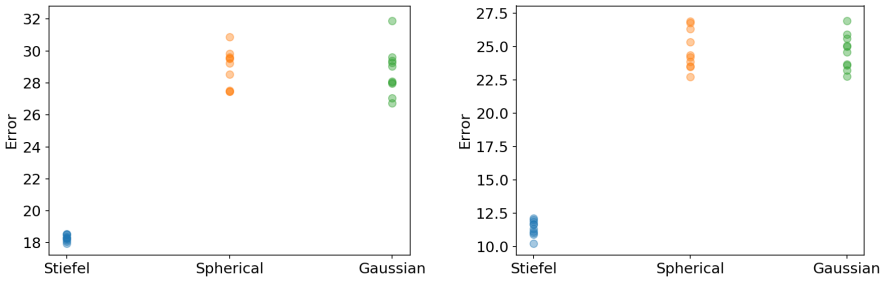
The estimator (Eq. 1) is also compared with the entry-wise estimator:

$$\widehat{\nabla} f^{\delta}_{E}(x) := [\widehat{\nabla}_i f^{\delta}_{E}(x)]_{i \in [n]},$$

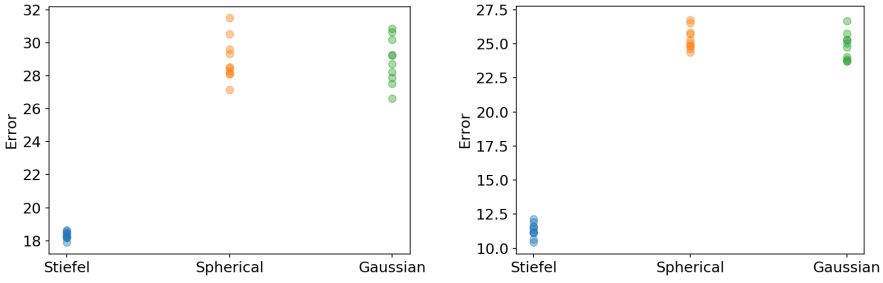
where $\widehat{\nabla}_i f^{\delta}_{E}(x) = \frac{1}{2\delta}\left(f(x + \delta \boldsymbol{e}_i) - f(x - \delta \boldsymbol{e}_i)\right)$ and $\boldsymbol{e}_i$ is the vector with 1 on the i-th entry and 0 on all other entries. The comparison results are summarized in Tables 1 and 2.



(a) $x = 0, \delta = 0.1, k = 300$ (b) $x = 0, \delta = 0.1, k = 400$

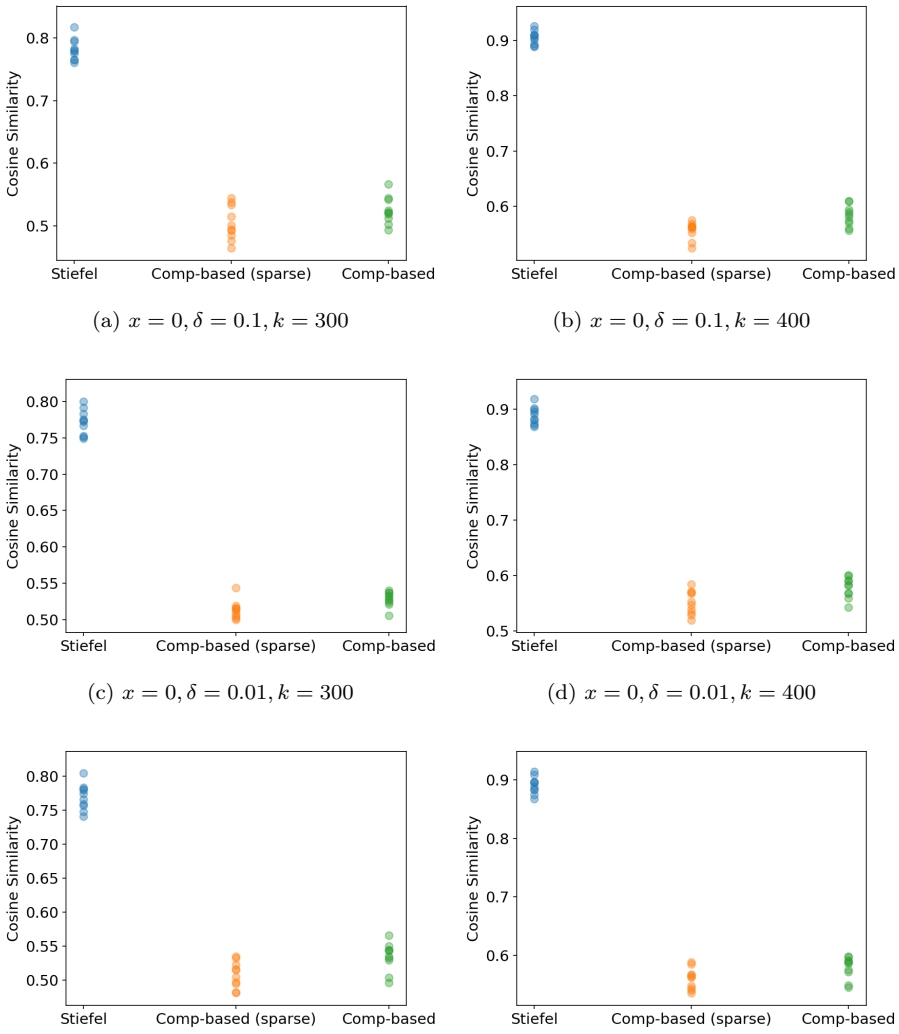


(c) $x = 0, \delta = 0.01, k = 300$ (d) $x = 0, \delta = 0.01, k = 400$



(e) $x = 0, \delta = 0.001, k = 300$ (f) $x = 0, \delta = 0.001, k = 400$

Figure 2: Errors of gradient estimators on the test function defined in (Eq. 7) with $n = 500$. Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. The x-axis labels the estimators: "Stiefel" is the estimator $\widehat{\nabla} f_k^\delta(x)$ (Eq. 1); "Spherical" is the estimator $\widehat{\nabla} f_{k,S}^\delta(x)$ (Eq. 5); "Gaussian" is the estimator $\widehat{\nabla} f_{k,G}^\delta(x)$ (Eq. 6). The y-axis is the error of the estimator. The error is $\left\|\widehat{\nabla} f_k^\delta(x) - \nabla f(x)\right\|$ (or $\left\|\widehat{\nabla} f_{k,S}^\delta(x) - \nabla f(x)\right\|$, $\left\|\widehat{\nabla} f_{k,G}^\delta(x) - \nabla f(x)\right\|$). Each dot represents one observed error of one estimator. Each estimator is evaluated 10 times (thus 10 dots for each estimator). For example, in subfigure (a), the 10 blue dots scattered above "Stiefel" show that the errors of 10 evaluations of $\widehat{f_k^\delta}(x)$ with parameters $x = 0, \delta = 0.1, k = 300$ are in range 17 to 20. More results for other values of (x, δ, k) can be found in the Appendix.



(a) $x = 0, \delta = 0.1, k = 300$ (b) $x = 0, \delta = 0.1, k = 400$

(c) $x = 0, \delta = 0.01, k = 300$ (d) $x = 0, \delta = 0.01, k = 400$

(e) $x = 0, \delta = 0.001, k = 300$ (f) $x = 0, \delta = 0.001, k = 400$

Figure 3: Performance of gradient estimators on the test function defined in (Eq. 7), measured by cosine similarity with the true gradient. Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. The x-axis labels the estimators: "Stiefel" is the estimator $\widehat{\nabla} f_k^\delta(x)$ (Eq. 1); "Comp-based (sparse)" is the estimator $\widehat{\nabla} f_{k,C}^\delta(x)$ (Algorithm 2) with sparsity parameter $s = 100$; "Comp-based" is the estimator $\widehat{\nabla} f_{k,C}^\delta(x)$ with sparsity parameter $s = \infty$ (no sparsity constrains). The y-axis is the cosine similarity between the estimator and the ground truth, which is $\frac{\left\langle \widehat{\nabla} f_k^\delta(x), \nabla f(x) \right\rangle}{\left\| \widehat{\nabla} f_k^\delta(x) \right\| \|\nabla f(x)\|}$ (or $\frac{\left\langle \widehat{\nabla} f_{k,C}^\delta(x), \nabla f(x) \right\rangle}{\left\| \widehat{\nabla} f_{k,C}^\delta(x) \right\| \|\nabla f(x)\|}$). Each dot represents one observed cosine similarity of one estimator. Each estimator is evaluated 10 times (thus 10 dots for each estimator). For example, in subfigure (a), the 10 blue dots scattered above "Stiefel" show that the cosine similarity of 10 evaluations of $\widehat{f}_k^\delta(x)$ with parameters $x = 0, \delta = 0.1, k = 300$ are in range 0.75 to 0.8. Note that larger cosine similarity means more accurate estimation. More results for other values of (x, δ, k) can be found in the Appendix.

Table 2: The errors of gradient estimators against finite-difference granularity δ. The first row shows δ; The second row shows errors of $\widehat{\nabla} f_n^\delta(x)$ ($n = 500$ is the dimension); The third row shows errors of the entry-wise estimator. The error of an estimator is its distance to the true gradient in Euclidean norm: $\left\|\widehat{\nabla} f_n^\delta(x) - \nabla f(x)\right\|$. Errors of $\widehat{\nabla} f_n^\delta(x)$ are in average $\pm$ standard derivation format, where each average and standard deviation gather information from 10 runs. The entry-wise estimator is not random and its error is computed in a single run. The test function is defined in (Eq. 7). For all evaluations in this table, $x = \frac{\pi}{4}\mathbf{1}$ is used.

δ	0.1	0.01	0.001
Stiefel's sampling errors	2.4e-4±1.0e-5	2.5e-6±1.5-07	2.5e-8±6.8e-10
Entry-wise errors	3.2e-2	3.2e-4	3.2e-6

Table 3: The errors of Hessian estimators against finite-difference granularity δ. The first row shows δ; The second row shows errors of $\widehat{\mathrm{H}} f_n^\delta(x)$ ($n = 100$ is the dimension); The third row shows errors of the entry-wise estimator $\widehat{\mathrm{H}} f_E^\delta(x)$. The error of an estimator is its distance to the true Hessian in spectral norm: $\left\|\widehat{\mathrm{H}} f_n^\delta(x) - \nabla^2 f(x)\right\|$ (or $\left\|\widehat{\mathrm{H}} f_E^\delta(x) - \nabla^2 f(x)\right\|$). Errors of $\widehat{\mathrm{H}} f_n^\delta(x)$ are in "average $\pm$ standard derivation" format, where each average and standard deviation gather information from 10 runs. The entry-wise estimator is not random and its error is computed in a single run. The test function is defined in (Eq. 7). For all evaluations in this table, $x = \frac{\pi}{2}\mathbf{1}$ is used.

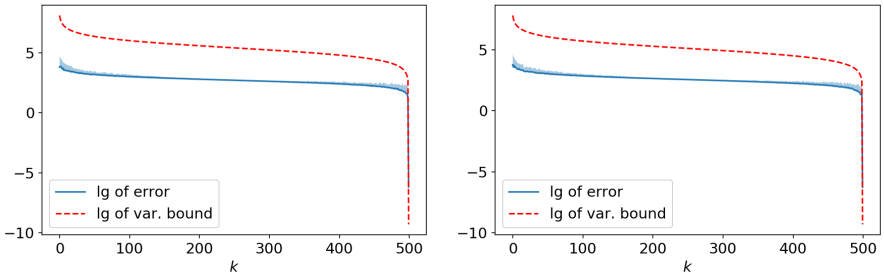
δ	0.1	0.01	0.001
Stiefel's sampling errors	0.17±0.024	1.7e-3±0.16e-4	1.6e-5±1.6e-6
Entry-wise errors	4.4	4.3e-2	4.3e-4

Table 4: The errors of Hessian estimators against finite-difference granularity δ. The first row shows δ; The second row shows errors of $\widehat{\mathrm{H}} f_n^\delta(x)$ ($n = 100$ is the dimension); The third row shows errors of the entry-wise estimator $\widehat{\mathrm{H}} f_E^\delta(x)$. The error of an estimator is its distance to the true Hessian in spectral norm: $\left\|\widehat{\mathrm{H}} f_n^\delta(x) - \nabla^2 f(x)\right\|$ (or $\left\|\widehat{\mathrm{H}} f_E^\delta(x) - \nabla^2 f(x)\right\|$). Errors of $\widehat{\mathrm{H}} f_n^\delta(x)$ are in "average $\pm$ standard derivation" format, where each average and standard deviation gather information from 10 runs. The entry-wise estimator is not random and its error is computed in a single run. The test function is defined in (Eq. 7). For all evaluations in this table, $x = \frac{\pi}{4}\mathbf{1}$ is used.

δ	0.1	0.01	0.001
Stiefel's sampling errors	4.1e-3±5.3e-4	3.8e-5±4.63e-6	3.8e-7±3.7e-8
Entry-wise errors	0.12	1.2e-3	1.2e-5

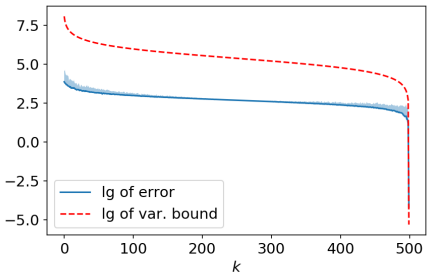
5.3 Empirical Verification of the Theorems

As discussed in the introduction (Figure 1), the error is highly aligned with the variance bound. Here we present more versions of Figure 1, with different values of δ and x. These results can be found in Figure 4.

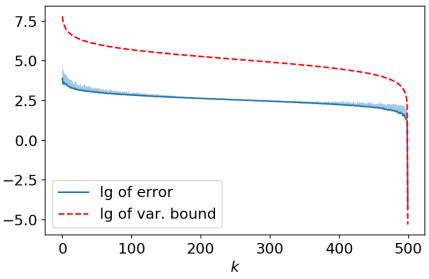


(a) $x = 0$, $\delta = 0.001$

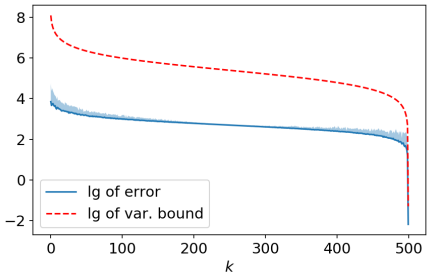
(b) $x = \frac{\pi}{4}\mathbf{1}$, $\delta = 0.001$



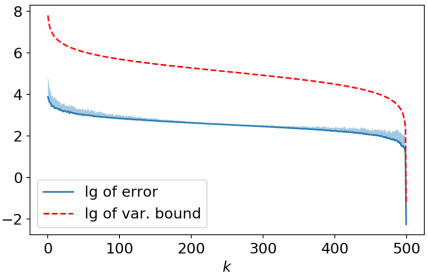
(c) $x = 0$, $\delta = 0.01$



(d) $x = \frac{\pi}{4}\mathbf{1}$, $\delta = 0.01$



(e) $x = 0$, $\delta = 0.1$



(f) $x = \frac{\pi}{4}\mathbf{1}$, $\delta = 0.1$

Figure 4: Errors of gradient estimators $\widehat{\nabla} f_k^\delta(x)$ with k ranging from 1 to n, in base-10 log-scale. Here $n = 500$. The underlying test function is $f(x) = \exp((x_1 - 1)(x_2 + 2)) + \sum_{j=1}^{500} \sin(x_j)$, where x_j denotes the j-th component of vector x. The location for estimation x, and the finite difference granularity δ are labeled in the captions of the subfigures. The solid blue curve plots the errors of gradient estimator, and is averaged over 10 runs. The shaded area above the solid curve shows 10 times standard deviation of the errors in logarithmic scale. The dashed red curve, as a function of k, is $c(k) = \lg\left(\|\nabla f(x)\|^2 \left(\frac{n}{k} - 1\right) + \delta^2 \left(\frac{n^2}{k} - n\right) \|\nabla f(x)\| + \frac{\delta^4 n^2}{k}\right)$, which is the base-10 log of of variance bound for the gradient estimators (up to constants).

6 Empirical Results for the Hessian Estimators

The Hessian estimators are empirically studied, in the same way that the gradient estimators are studied. Similar to the gradient case, (i) the errors of the Hessian estimators also draw an "S"-shape curve in logarithmic scale (See Figure 6); (ii) When $k = n$, the stochastic Hessian estimator (Eq. 2) outperform the classic entry-by-entry Hessian estimator (See Tables 3 and 4). When k is much smaller than n, the supremacy of Hessian estimator (Eq. 2) is sometimes less pronounced. In particular, the estimator (Eq. 2) may have same level of accuracy as the estimator introduced by Wang (2022) (See Figure 14 in the Appendix for details).

6.1 Comparison with Stochastic Estimators

For the same number of random direction specified by k, and finite difference granularity δ, we compare the estimators $\widehat{\mathrm{H}} f_k^\delta(x)$ with:

- The estimator via spherical sampling (Wang, 2022):

$$\widehat{\mathrm{H}} f_{k,S}^\delta(x) := \frac{n^2}{8k\delta^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \Big(f(x + \delta v_i + \delta w_j) - f(x + \delta v_i - \delta w_j) \\ - f(x - \delta v_i + \delta w_j) + f(x - \delta v_i - \delta w_j) \Big) \left(v_i w_j^\top + w_j v_i^\top \right), \quad (8)$$

  where $v_1, v_2, \cdots, v_k$ and $w_1, w_2, \cdots, w_k$ are *i.i.d.* sampled from the uniform distribution over $\mathbb{S}^{n-1}$.

- The estimator via Gaussian sampling and the Stein's identity (Balasubramanian and Ghadimi, 2021):

$$\widehat{\mathrm{H}} f_{k,G}^\delta(x) := \frac{n}{2k^2\delta^2} \sum_{i=1}^{k^2} \left(f\left(x + \frac{\delta v_i}{\sqrt{n}} \right) - 2f(x) + f\left(x - \frac{\delta v_i}{\sqrt{n}} \right) \right) (v_i v_i - I), \quad (9)$$

  where $v_1, v_2, \cdots, v_k \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$. Note that the finite difference step size is downscale by a factor of $\sqrt{n}$ so that the expected granularity is of order $\Theta(\delta)$.

All methods are tested using the test function (Eq. 7) with dimension $n = 100$. All function evaluations are *noise-free*.

6.2 Comparison with the Entry-wise Estimator

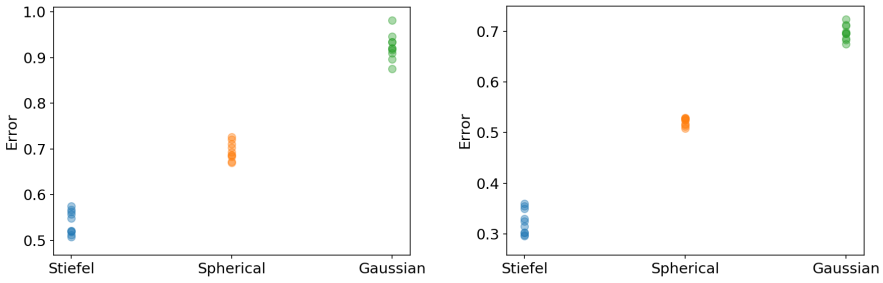
The estimator (Eq. 2) is also compared with the entry-wise estimator:

$$\widehat{\mathrm{H}} f_E^\delta(x) := [\widehat{\mathrm{H}}_{ij} f_E^\delta(x)]_{i,j \in [n]},$$

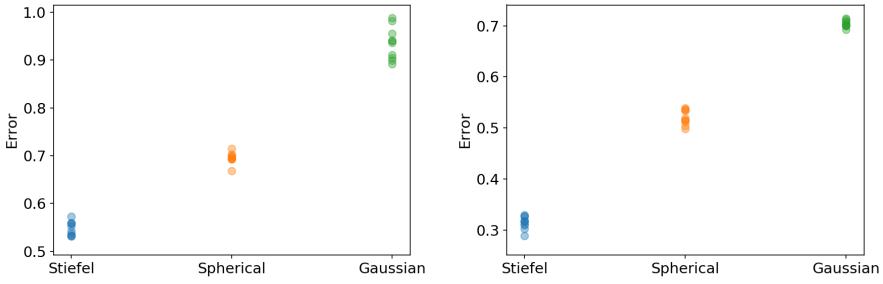
where $\widehat{\mathrm{H}}_{ij} f_E^\delta(x) = \frac{1}{4\delta^2} \left(f(x + \delta \boldsymbol{e}_i + \delta \boldsymbol{e}_j) - f(x - \delta \boldsymbol{e}_i + \delta \boldsymbol{e}_j) - f(x + \delta \boldsymbol{e}_i - \delta \boldsymbol{e}_j) + f(x - \delta \boldsymbol{e}_i - \delta \boldsymbol{e}_j) \right)$ and $\boldsymbol{e}_i$ is the vector with 1 on the i-th entry and 0 on all other entries. The comparison results are summarized in Tables 3 and 4.

6.3 Empirical Verification of the Theorems

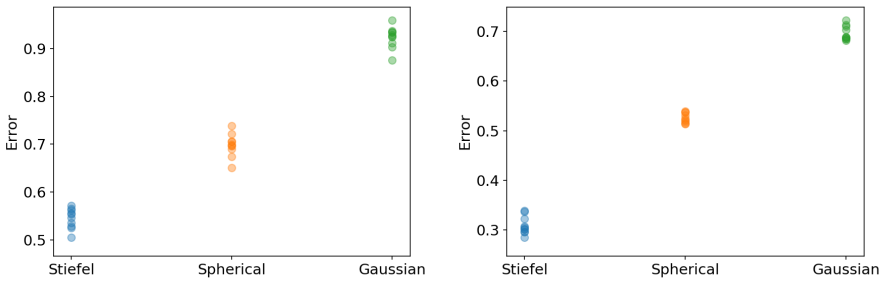
As discussed in the introduction (Figure 1), the error is highly aligned with the variance bound. Here we present the Hessian counterpart of Figure 1, with different values of δ and x.



(a) $x = 0, \delta = 0.1, k = 60$ (b) $x = 0, \delta = 0.1, k = 80$

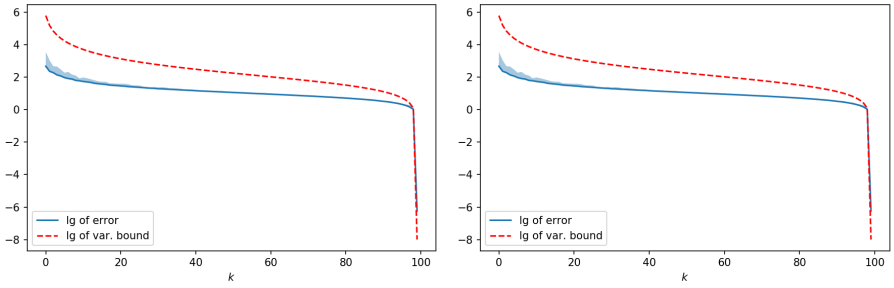


(c) $x = 0, \delta = 0.01, k = 60$ (d) $x = 0, \delta = 0.01, k = 80$

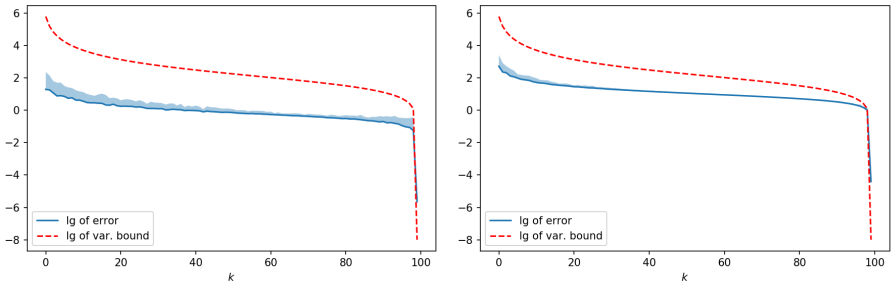


(e) $x = 0, \delta = 0.001, k = 60$ (f) $x = 0, \delta = 0.001, k = 80$

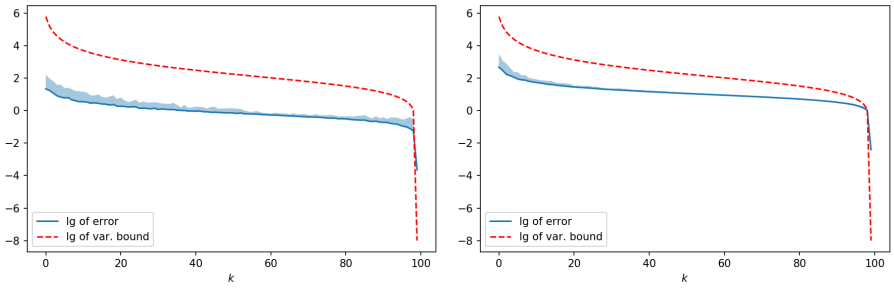
Figure 5: Errors of Hessian estimators on the test function defined in (Eq. 7) with $n = 100$. Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. The x-axis labels the estimators: "Stiefel" is the estimator $\widehat{\mathrm{H}} f_k^\delta(x)$ (Eq. 2); "Spherical" is the estimator $\widehat{\mathrm{H}} f_{k,S}^\delta(x)$ (Eq. 8); "Gaussian" is the estimator $\widehat{\mathrm{H}} f_{k,G}^\delta(x)$ (Eq. 9). The y-axis is the error of the estimator. The error is $\left\|\widehat{\mathrm{H}} f_k^\delta(x) - \nabla^2 f(x)\right\|_F$ (or $\left\|\widehat{\mathrm{H}} f_{k,S}^\delta(x) - \nabla^2 f(x)\right\|_F$, $\left\|\widehat{\mathrm{H}} f_{k,G}^\delta(x) - \nabla^2 f(x)\right\|_F$). Each dot represents one observed error of one estimator. Each estimator is evaluated 10 times (thus 10 dots for each estimator). For example, in subfigure (a), the 10 blue dots scattered above "Stiefel" show the errors of 10 evaluations of $\widehat{\mathrm{H}} f_k^\delta(x)$ with parameters $x = 0, \delta = 0.1, k = 100$. More results for other values of (x, δ, k) can be found in Appendix B.



(a) $x = \frac{\pi}{4}$, $\delta = 0.001$ (b) $x = \frac{\pi}{2}\mathbf{1}$, $\delta = 0.001$



(c) $x = \frac{\pi}{4}$, $\delta = 0.01$ (d) $x = \frac{\pi}{2}\mathbf{1}$, $\delta = 0.01$



(e) $x = \frac{\pi}{4}$, $\delta = 0.1$ (f) $x = \frac{\pi}{2}\mathbf{1}$, $\delta = 0.1$

Figure 6: Errors of Hessian estimators $\widehat{\mathrm{H}}f_k^{\delta}(x)$ as k ranges from 1 to n, in base-10 log-scale. Here $n = 100$. The underlying test function is $f(x) = \exp((x_1-1)(x_2+2)) + \sum_{j=1}^{100} \sin(x_j)$, where x_j denotes the j-th component of vector x. The location for estimation x, and the finite difference granularity δ are labeled in the captions of the subfigures. The solid blue curve plots errors of the Hessian estimators (in terms of spectral norm), and is averaged over 10 runs. The shaded area above the solid curve shows 10 times standard deviation of the errors in logarithmic scale. The dashed red curve, as a function of k, is $c(k) = \lg\left(\left\|\nabla^2 f(x)\right\|_F^2 \left(\frac{n^2}{k^2} - 1\right) + 2\delta^2 \left\|\nabla^2 f(x)\right\| \left(\frac{n^4}{k^2} - n^2\right) + \left\|\nabla^2 f(x)\right\| \frac{n^4\delta^4}{k^2}\right)$, which is the base-10 log of of variance bound for the Hessian estimators (up to constants).

7 Conclusion

We introduce gradient and Hessian estimators using random orthogonal frames sampled from the Stiefel's manifold. The methods extend previous gradient/Hessian estimators based on spherical sampling (Flaxman et al., 2005; Wang, 2022). Theoretically and empirically, we show that the variance of the estimation is reduced, and the accuracy of the estimation is improved. Refined bias bounds via Taylor expansion of higher orders are also provided.

References

Balasubramanian, K. and Ghadimi, S. (2021). Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pages 1–42.

Cai, H., McKenzie, D., Yin, W., and Zhang, Z. (2022). A one-bit, comparison-based gradient estimator. *Applied and Computational Harmonic Analysis*, 60:242–266.

Chikuse, Y. (2003). *Statistics on Special Manifolds*. Springer New York, NY.

Conn, A. R., Scheinberg, K., and Vicente, L. N. (2009). *Introduction to derivative-free optimization*. SIAM.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2015). Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning.

Jamieson, K. G., Nowak, R., and Recht, B. (2012). Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems*, 25.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205.

Nesterov, Y. and Spokoiny, V. (2017). Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566.

Plan, Y. and Vershynin, R. (2012). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494.

Plan, Y. and Vershynin, R. (2014). Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461.

Raginsky, M. and Rakhlin, A. (2011). Information-based complexity, feedback and dynamics in convex programming. *IEEE Transactions on Information Theory*, 57(10):7036–7056.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Stein, C. M. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135 – 1151.

Wang, T. (2022). Towards Sharp Stochastic Zeroth Order Hessian Estimators over Riemannian Manifolds. *arXiv preprint arXiv:2201.10780*.

Wang, T., Huang, Y., and Li, D. (2021). From the Greene–Wu Convolution to Gradient Estimation over Riemannian Manifolds. *arXiv preprint arXiv:2108.07406*.

A Proofs

A.1 Proof of Proposition 1

Proof. For any $\tau > 0$ and $x, v \in \mathbb{R}^n$ with $\|v\| = 1$, define $x_{\tau,v} = x + \tau v$. When τ is small, Taylor's theorem and (k, L)-smoothness of f give

$$\left|\frac{\partial^{k-1} f(x_{\tau,v})[v] - \partial^{k-1} f(x)[v]}{\tau}\right| \overset{①}{=} \left|\partial^k f(z_{\tau,v})\,[v]\right|,$$

where $z_{\tau,v}$ depends on τ and v and $\lim_{\tau\to 0} z_{\tau,v} = x$ for any v.

Since the f is (k, L)-smooth, for any $v \in \mathbb{S}^{n-1}$ (the unit sphere in $\mathbb{R}^n$), it holds that

$$\left|\frac{\partial^{k-1} f(x_{\tau,v})[v] - \partial^{k-1} f(x)[v]}{\tau}\right| \overset{②}{\leq} L.$$

Combining ① and ② gives

$$\left|\partial^k f(z_{\tau,v})\,[v]\right| \leq L,$$

for any $v \in \mathbb{S}^{n-1}$ and any sufficiently small τ. Thus for any $x \in \mathbb{R}^n$, we have

$$L \geq \sup_{v\in\mathbb{S}^{n-1}} \lim_{\tau\to 0} \left|\partial^k f(z_{\tau,v})\,[v]\right| = \left\|\partial^k f(x)\right\|.$$

□

A.2 Proofs of Propositions 2, 3 and 4

We first prove the following proposition, which will be useful in proving Proposition 3.

Proposition 6. *Let v be uniformly sampled from $\mathbb{S}^{n-1}$ ($n \geq 2$). It holds that*

$$\mathbb{E}\,[v_i^p] = \frac{(p-1)(p-3)\cdots 1}{n(n+2)\cdots(n+p-2)}$$

for all $i = 1, 2, \cdots, n$ and any positive even integer p.

Proof. Let $(r, \varphi_1, \varphi_2, \cdots, \varphi_{n-1})$ be the spherical coordinate system. We have, for any $i = 1, 2, \cdots, n$ and an even integer p,

$$\mathbb{E}\,[v_1^p] = \frac{1}{A_n}\int_0^{2\pi}\int_0^{\pi}\cdots\int_0^{\pi} \cos^p(\varphi_1)\sin^{n-2}(\varphi_1)\sin^{n-3}(\varphi_2)\cdots\sin(\varphi_{n-2})\,d\varphi_1\,d\varphi_2\cdots d\varphi_{n-1},$$

where A_n is the surface area of $\mathbb{S}^{n-1}$. Let

$$I(n,p) := \int_0^{\pi} \sin^n(x)\cos^p(x)\,dx.$$

Clearly, $I(n,p) = I(n,p-2) - I(n+2,p-2)$. By integration by parts, we have $I(n+2,p-2) = \frac{n+1}{p-1} I(n,p)$. The above two equations give $I(n,p) = \frac{p-1}{n+p} I(n,p-2)$.

Thus we have $\mathbb{E}\,[v_1^p] = \frac{I(n-2,p)}{I(n-2,0)} = \frac{I(n-2,p)}{I(n-2,p-2)}\frac{I(n-2,p-2)}{I(n-2,p-4)}\cdots\frac{I(n-2,2)}{I(n-2,0)} = \frac{(p-1)(p-3)\cdots 1}{n(n+2)\cdots(n+p-2)}$. We conclude the proof by by symmetry.

□

Proof of Proposition 2. Let $p = 2$ in Proposition 6, we have $\mathbb{E}\left[v_i^2\right] = \frac{1}{n}$ for all i. In addition, when $i \neq j$, $\mathbb{E}\left[v_i v_j | v_j = a\right] = 0$ for any a. Thus $\mathbb{E}\left[v_i v_j\right] = 0$ for all $i \neq j$. □

Proof of Proposition 3. By Proposition 6, we have

$$\mathbb{E}\left[v_i^4\right] = \frac{3}{n(n+2)}.$$

By symmetry and that $\left(\sum_{i=1}^n v_i^2\right)^2 = 1$, for any $i \neq j$, we have

$$\mathbb{E}\left[v_i^2 v_j^2\right] = \frac{1 - \frac{3n}{n(n+2)}}{n(n-1)} = \frac{1}{n(n+2)}.$$

In addition, similar to the proof for $\mathbb{E}\left[v_i v_j\right] = 0$ for $i \neq j$ (Proposition 2), we have $\mathbb{E}\left[v_i v_j v_k v_l\right] = 0$ for any $i, j, k, l \in \{1, 2, \cdots, n\}$ and $i \notin \{j, k, l\}$. □

A.3 Proof of Proposition 4

Proof. Let δ_k^l be the Kronecker delta. Using Einstein's notation, Proposition 2 is equivalent to $\mathbb{E}\left[v_i v^j\right] = \frac{1}{n}\delta_i^j$. Thus we have

$$\mathbb{E}\left[\left(v^\top A w\right) v w^\top\right] = \mathbb{E}\left[v_k A_l^k w^l v^i w_j\right] = \frac{1}{n^2} A_l^k \delta_k^i \delta_j^l = \frac{1}{n^2} A_j^i,$$

which concludes the proof.

□

A.4 Proof of Theorem 2(a)

The original proof was due to Flaxman et al. (2005). Here we present a proof via the divergence theorem. This version of proof will also assist the proof for Theorem 4(a). Define

$$f^\delta(x) = \frac{1}{\delta^n V_n} \int_{\mathbb{B}^n} f(x + \delta v)\, dv,$$

where V_n is the volume of $\mathbb{B}^n$.

Lemma 1. *Let f be a smooth function. For any $k \in \mathbb{N}_+$ and $v_1, v_2, \cdots, v_k \in \mathbb{R}^n$ sampled from the Gram-Schmidt sampling process, it holds that*

$$\frac{n}{\delta}\mathbb{E}\left[\nabla \widehat{f}_k^\delta(x)\right] = \nabla f^\delta(x),$$

for all $x \in \mathbb{R}^n$, and all $k = 1, 2, \cdots, n$.

Proof. Without loss of generality, let $x = 0$. Derivations for other values of x follows similar arguments. Let u be an arbitrary unit vector in $\mathbb{R}^n$, and let U be the constant vector field generated by u. Let $X = fU$, which is the vector field U multiplied by the function values of f. Apply the divergence theorem to this vector field and the region enclosed by $\delta\mathbb{S}^{n-1}$ gives

$$\int_{\delta\mathbb{B}^n} \nabla \cdot X\, dV = \int_{\delta\mathbb{S}^{n-1}} X \cdot d\vec{S}.$$

The above equation is equivalent to

$$\left\langle \int_{v \in \delta\mathbb{B}^n} \nabla f(v)\, dv, u \right\rangle = \left\langle \int_{v \in \delta\mathbb{S}^{n-1}} f(v) \frac{v}{\|v\|}\, dv, u \right\rangle, \quad \forall x \in \mathbb{R}^n$$

By the dominated convergence theorem, we can exchange the gradient and the integral to get

$$\left\langle \nabla \int_{v \in \delta\mathbb{B}^n} f(v)\, dv, u \right\rangle = \left\langle \int_{v \in \delta\mathbb{S}^{n-1}} f(v) \frac{v}{\|v\|}\, dv, u \right\rangle, \quad \forall u \in \mathbb{R}^n,$$

or equivalently

$$\left\langle \delta^n V_n \nabla f^\delta(0), u \right\rangle \overset{①}{=} \left\langle \int_{v\in\mathbb{S}^{n-1}} f(\delta v) v \, dv, u \right\rangle, \quad \forall u \in \mathbb{R}^n,$$

where on the right-hand-side a change of integration from $\delta\mathbb{S}^{n-1}$ to $\mathbb{S}^{n-1}$ is used.

By Proposition 5, it holds that, for any v_i generated from the Gram-Schmidt sampling process,

$$\mathbb{E}\left[f(\delta v_i) v_i\right] \overset{②}{=} \frac{1}{\delta^{n-1} A_n} \int_{v_i\in\mathbb{S}^{n-1}} f(\delta v_i) v_i \, dv_i.$$

Combining ① and ② gives

$$\left\langle \frac{n}{\delta} \mathbb{E}\left[f(\delta v_i) v_i\right], u \right\rangle = \left\langle \nabla f^\delta(0), u \right\rangle.$$

The above equation concludes the proof since u is an arbitrary (unit) vector in $\mathbb{R}^n$. □

Proof of Theorem 2(a). Let V_n be the volume of $\mathbb{B}^n$, and let A_n be the area of $\mathbb{S}^{n-1}$. If f is $(1, L_1)$-smooth, we have

$$\left\|\nabla f^\delta(x) - \nabla f(x)\right\| \le \int_{\delta\mathbb{B}^n} \frac{1}{\delta^n V_n} \left\|\nabla f^\delta(x+v) - \nabla f(x)\right\| dv \le \int_0^\delta \frac{1}{\delta^n V_n} L_1 r^n A_n \, dr = \frac{L_1 n \delta}{n+1},$$

where the last equation uses $A_n = nV_n$. Combine the above result with Lemma 1 finishes the proof. □

A.5 Proof of Theorem 4(a)

Define

$$\widetilde{f}^\delta(x) = \frac{1}{\delta^n V_n} \frac{1}{\delta^n V_n} \int_{\mathbb{B}^n} \int_{\mathbb{B}^n} f(x + \delta v + \delta w) \, dw \, dv,$$

where V_n is the volume of $\mathbb{B}^n$.

Lemma 2. *Let f be twice continuously differentiable. It holds that*

$$\frac{n^2}{\delta^2} \mathbb{E}\left[\widehat{\mathrm{H}} f_k^\delta(x)\right] = \nabla^2 \widetilde{f}^\delta(x),$$

for any $x \in \mathbb{R}^n$ and $k = 1, 2, \cdots, n$.

Proof. Without loss of generality, we consider $x = 0$. Also, by linearity of expectation, it suffices to prove that for any v, w independently uniformly sampled from $\mathbb{S}^{n-1}$, we have

$$\mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}^{n-1}} \left[f(\delta v + \delta w) v w^\top\right] = \nabla^2 \widetilde{f}^\delta(0).$$

Let z be an arbitrary unit vector in $\mathbb{R}^n$. Let $g_z(x) = \nabla_z f^\delta(x) = \left\langle \nabla f^\delta(x), z \right\rangle$, where f^δ is defined in Appendix A.4. Let u be an arbitrary unit vector in $\mathbb{R}^n$, and let U be the constant vector field generated by u. Let $X_z := g_z U$ be the vector field U multiplied by the function values of g_z. Apply the divergence theorem to this vector field gives

$$\int_{\delta\mathbb{B}^n} \nabla \cdot X_z \, dV = \int_{\delta\mathbb{S}^{n-1}} X_z \cdot d\vec{S}.$$

The above equation is equivalent to

$$\left\langle \int_{\delta\mathbb{B}^n} \nabla g_z(v) \, dv, u \right\rangle = \left\langle \int_{\delta\mathbb{S}^{n-1}} g_z(v) \frac{v}{\|v\|} \, dv, u \right\rangle.$$

By the dominated convergence theorem, we can exchange the gradient and the integral to get

$$\left\langle \nabla \int_{\delta\mathbb{B}^n} g_z(v)\, dv, u \right\rangle = \left\langle \int_{\delta\mathbb{S}^{n-1}} g_z(v) \frac{v}{\|v\|}\, dv, u \right\rangle, \quad \forall u \in \mathbb{R}^n,$$

which is equivalent to

$$\left\langle \nabla \int_{\delta\mathbb{B}^n} \left\langle \nabla f^\delta(v), z \right\rangle dv, u \right\rangle \overset{①}{=} \left\langle \int_{\delta\mathbb{S}^{n-1}} \left\langle \nabla f^\delta(v), z \right\rangle \frac{v}{\|v\|}\, dv, u \right\rangle, \quad \forall u \in \mathbb{R}^n.$$

By Lemma 1, the right-hand-side of ① is

$$\begin{aligned}
&\left\langle \int_{\delta\mathbb{S}^{n-1}} \left\langle \nabla f^\delta(v), z \right\rangle \frac{v}{\|v\|}\, dv, u \right\rangle \\
&= \left\langle \int_{\mathbb{S}^{n-1}} \left\langle \frac{n}{\delta} \int_{\mathbb{S}^{n-1}} f(\delta v + \delta w) w\, dw, z \right\rangle v\, dv, u \right\rangle \\
&= \frac{n}{\delta} \delta^{n-1} A_n u^\top \mathbb{E}_{v,w} \left[f(\delta v + \delta w) v w^\top \right] z,
\end{aligned}$$

where A_n is the surface area of $\mathbb{S}^{n-1}$.

By dominated convergence theorem, we can interchange the integral and the directional derivative. Thus the left-hand-side of ① is

$$\left\langle \nabla \int_{\delta\mathbb{B}^n} \left\langle \nabla f^\delta(v), z \right\rangle dv, u \right\rangle = \delta^n V_n u^\top \nabla^2 \widetilde{f}^\delta(0) z,$$

where V_n is the volume of $\mathbb{B}^n$.

Since $A_n = nV_n$, collecting terms gives

$$u^\top \left(\nabla^2 \widetilde{f}^\delta(0) \right) z = \frac{n^2}{\delta^2} u^\top \mathbb{E}_{v,w} \left[f(\delta v + \delta w) v w^\top \right] z,$$

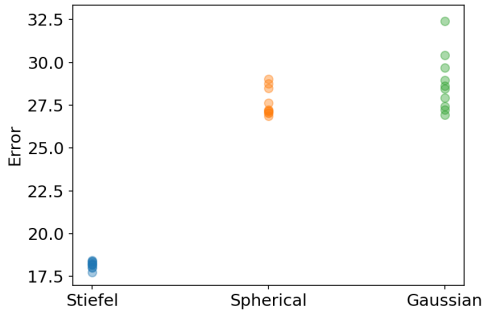
We conclude the proof by noting that the above is true for any (unit) vectors u and z. □

Proof of Theorem 4(a). Let V_n be the volume of $\mathbb{B}^n$, and let A_n be the area of $\mathbb{S}^{n-1}$. If f is $(2, L_2)$-smooth, we have

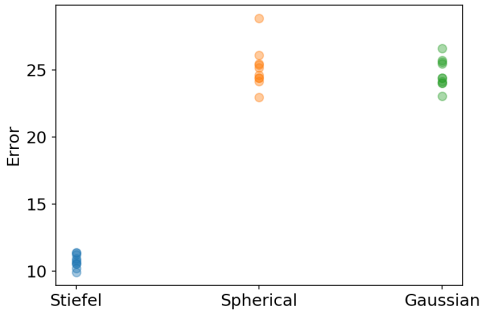
$$\begin{aligned}
\left\| \nabla^2 \widetilde{f}^\delta(x) - \nabla^2 f(x) \right\| &\le \int_{\delta\mathbb{B}^n} \int_{\delta\mathbb{B}^n} \frac{1}{\delta^{2n} V_n^2} \left\| \nabla^2 \widetilde{f}^\delta(x + v + w) - \nabla^2 f(x) \right\| dw\, dv \\
&\le \int_0^\delta \int_0^\delta \frac{1}{\delta^{2n} V_n^2} L_2 (r + s) r^{n-1} s^{n-1} A_n^2\, dr\, ds \\
&= \frac{2nL_2\delta}{n+1},
\end{aligned}$$

where the last equation uses $A_n = nV_n$. Combine the above result with Lemma 2 finishes the proof. □

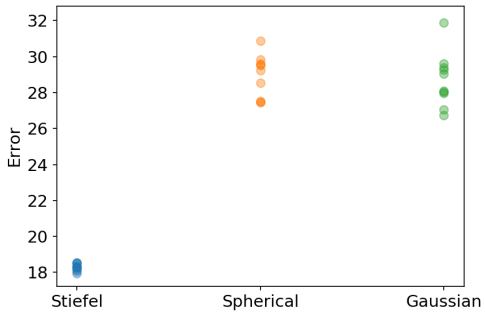
B Supplementary Figures



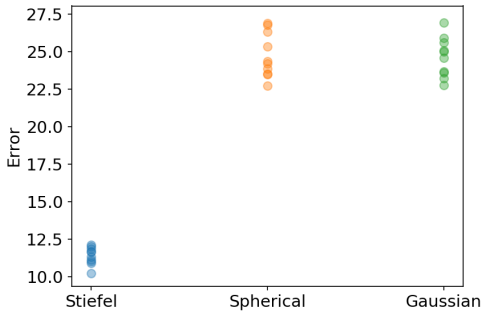
(a) $x = 0, \delta = 0.1, k = 300$



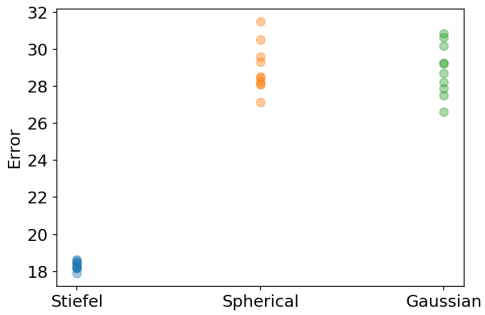
(b) $x = 0, \delta = 0.1, k = 400$



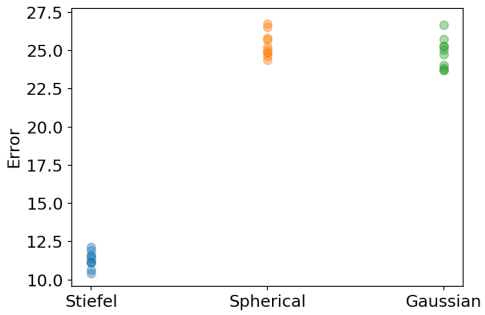
(c) $x = 0, \delta = 0.01, k = 300$



(d) $x = 0, \delta = 0.01, k = 400$

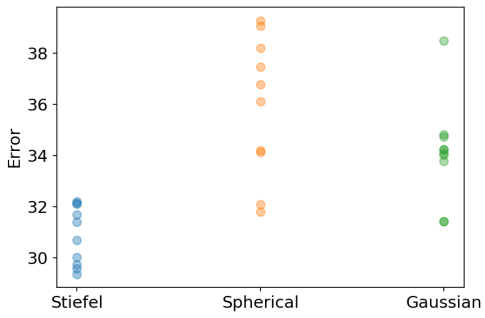


(e) $x = 0, \delta = 0.001, k = 300$

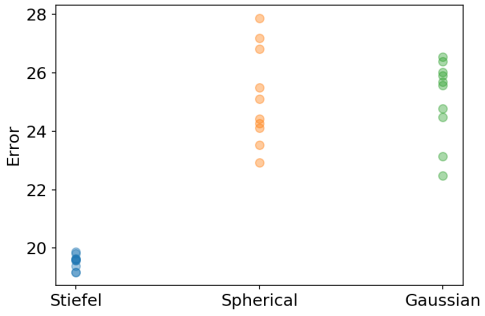


(f) $x = 0, \delta = 0.001, k = 400$

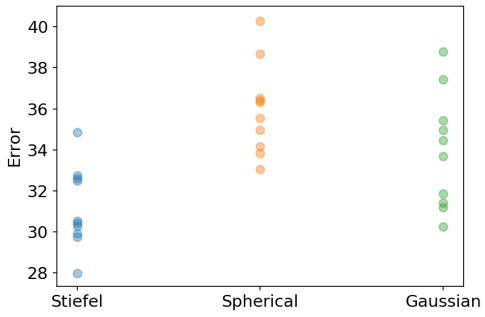
Figure 7: Errors of $\widehat{\nabla} f_k^\delta(x)$, $\widehat{\nabla} f_{k,B}^\delta(x)$, $\widehat{\nabla} f_{k,G}^\delta(x)$ on test function (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See caption of Figure 2 for detailed illustration.



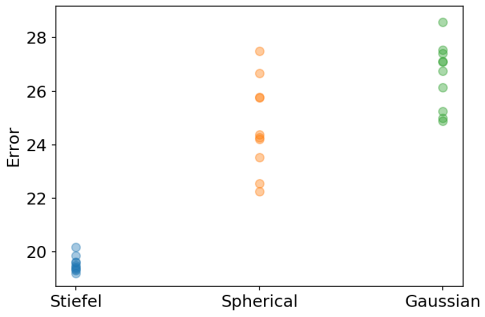
(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 100$



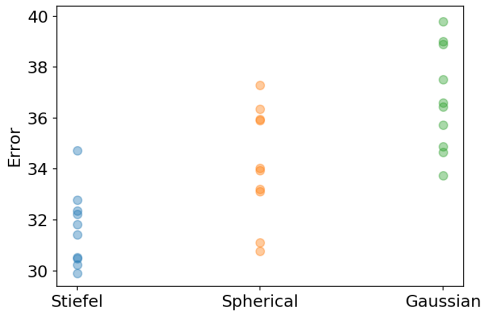
(b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 200$



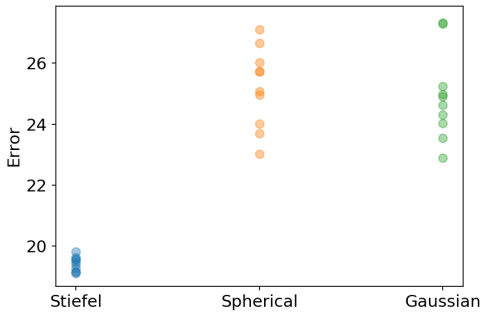
(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 100$



(d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 200$

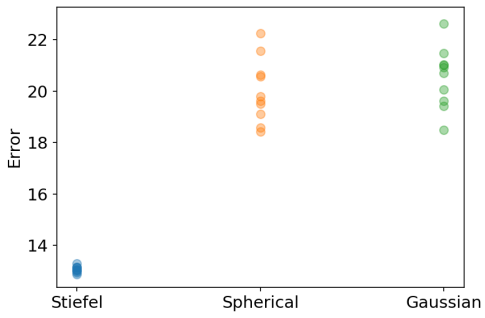


(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 100$

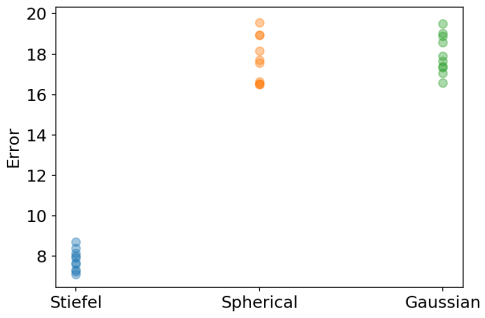


(f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 200$

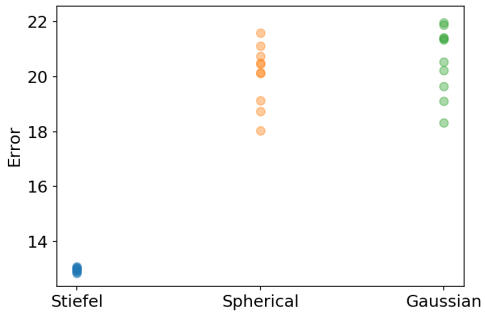
Figure 8: Errors of $\widehat{\nabla} f_k^\delta(x)$, $\widehat{\nabla} f_{k,B}^\delta(x)$, $\widehat{\nabla} f_{k,G}^\delta(x)$ on test function (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See caption of Figure 2 for detailed illustration.



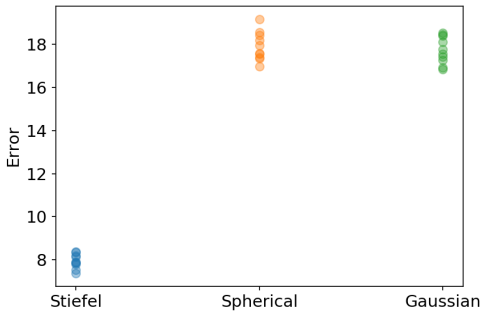
(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 300$



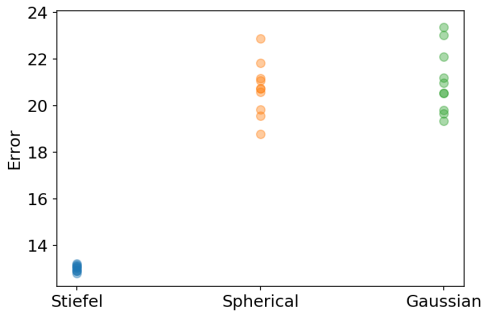
(b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 400$



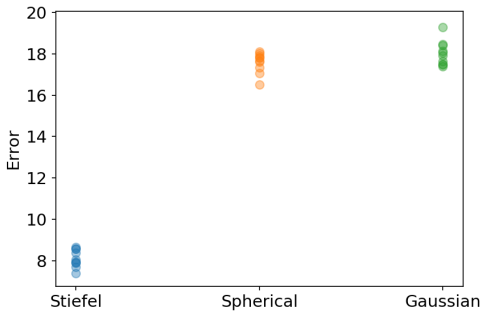
(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 300$



(d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 400$

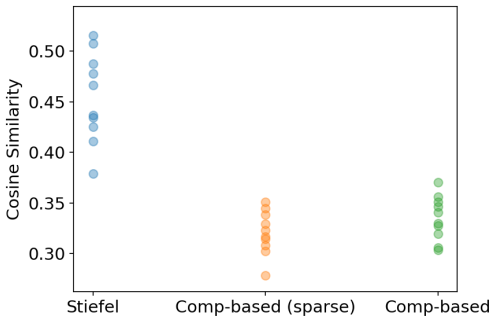


(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 300$

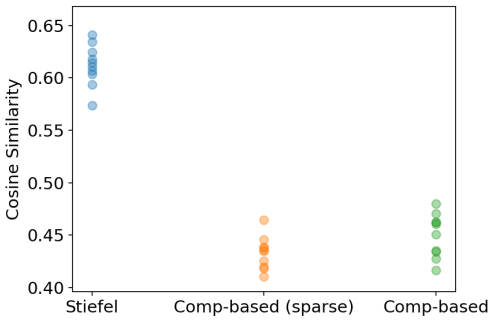


(f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 400$

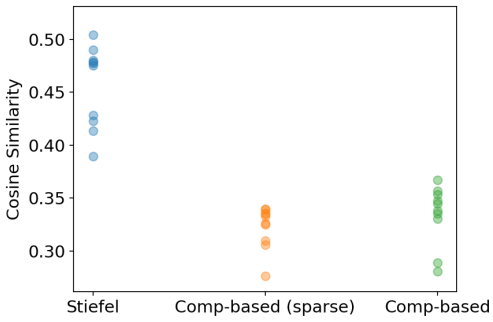
Figure 9: Errors of $\widehat{\nabla} f_k^\delta(x)$, $\widehat{\nabla} f_{k,B}^\delta(x)$, $\widehat{\nabla} f_{k,G}^\delta(x)$ on test function (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See caption of Figure 2 for detailed illustration.



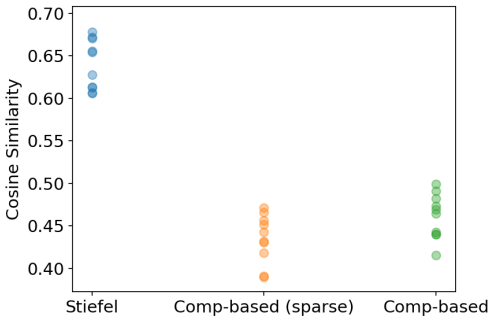
(a) $x = 0, \delta = 0.1, k = 100$



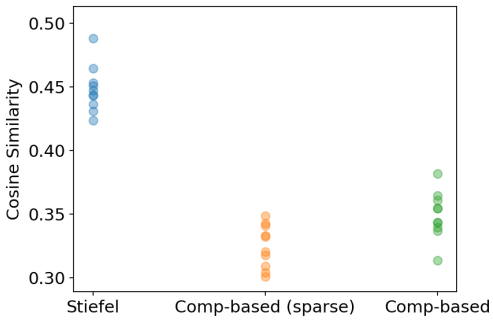
(b) $x = 0, \delta = 0.1, k = 200$



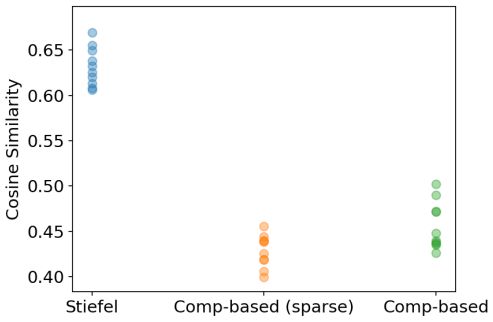
(c) $x = 0, \delta = 0.01, k = 100$



(d) $x = 0, \delta = 0.01, k = 200$

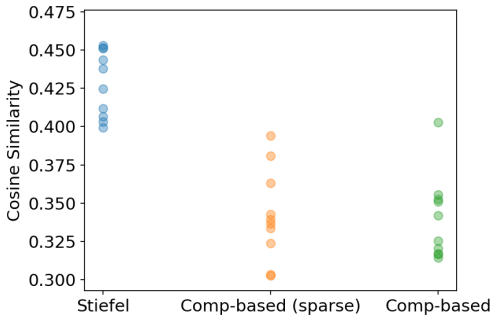


(e) $x = 0, \delta = 0.001, k = 100$

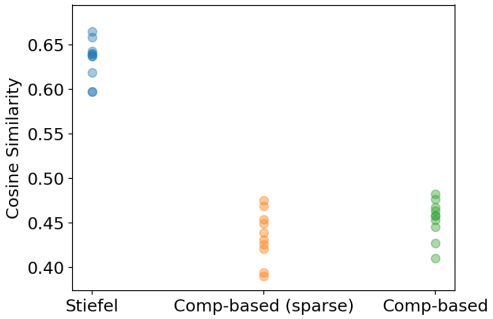


(f) $x = 0, \delta = 0.001, k = 200$

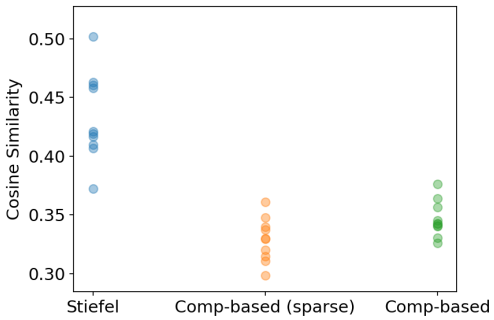
Figure 10: Cosine similarity between the gradient estimation and the true gradient, using the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See the caption of Figure 3 for detailed illustration.



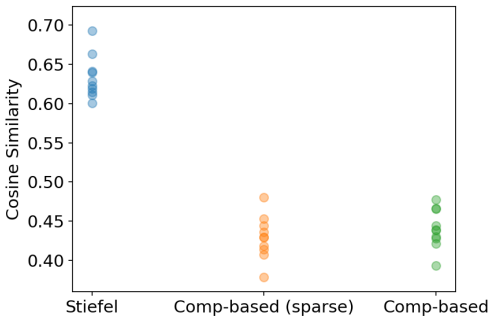
(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 100$



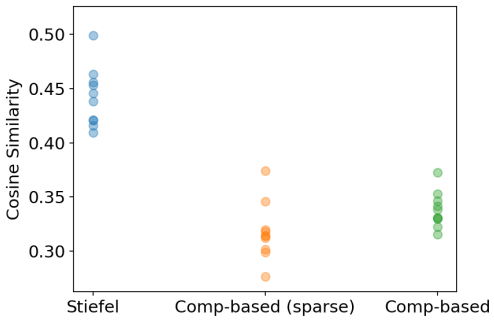
(b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 200$



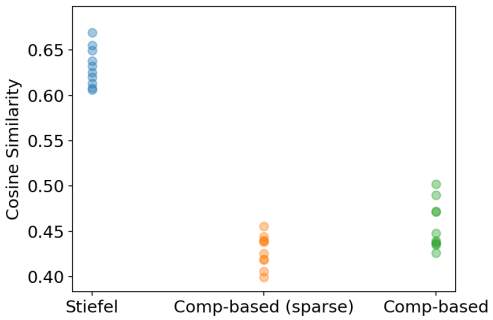
(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 100$



(d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 200$

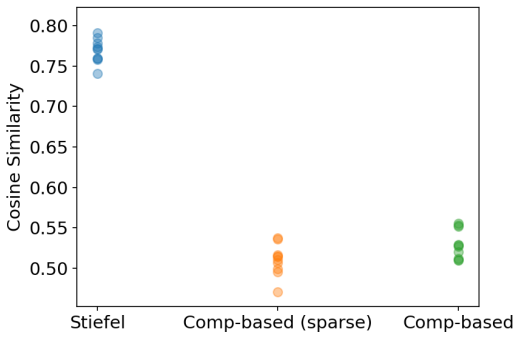


(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 100$

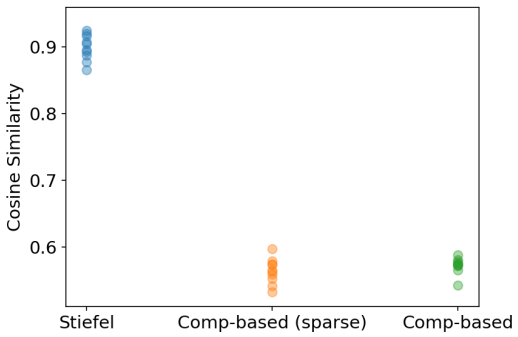


(f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 200$

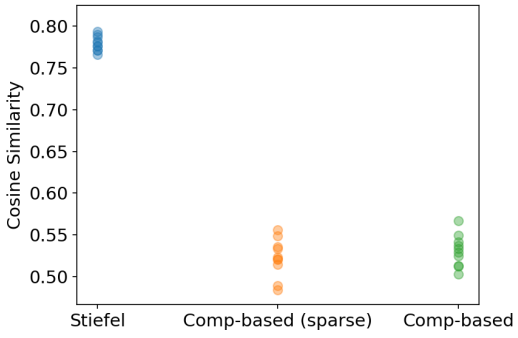
Figure 11: Cosine similarity between the gradient estimation and the true gradient, using the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See the caption of Figure 3 for detailed illustration.



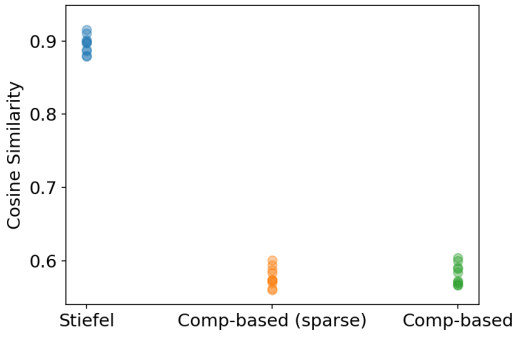
(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 300$



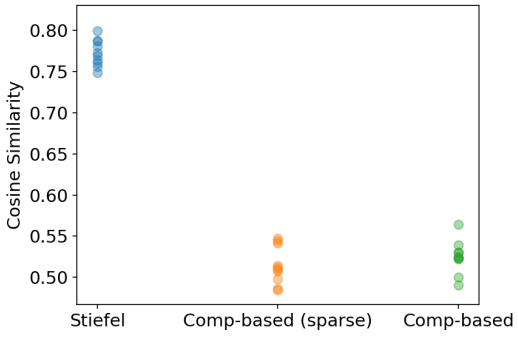
(b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 400$



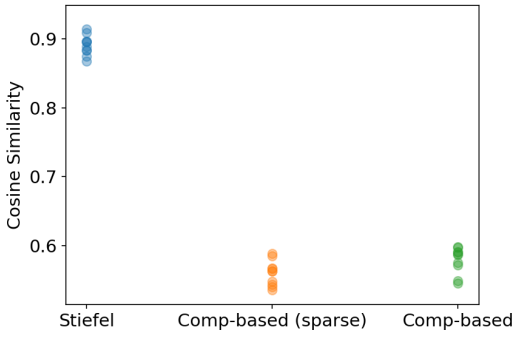
(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 300$



(d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 400$

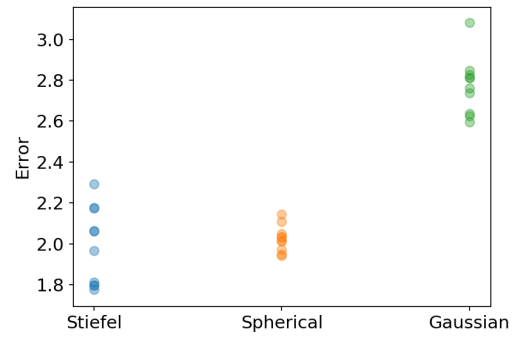


(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 300$

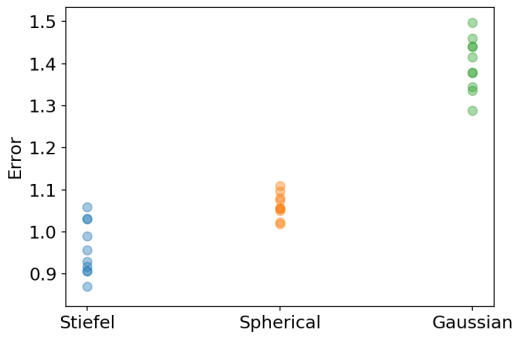


(f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 400$

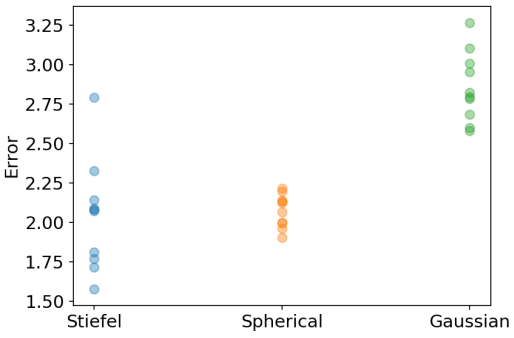
Figure 12: Cosine similarity between the gradient estimation and the true gradient, using the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See the caption of Figure 3 for detailed illustration.



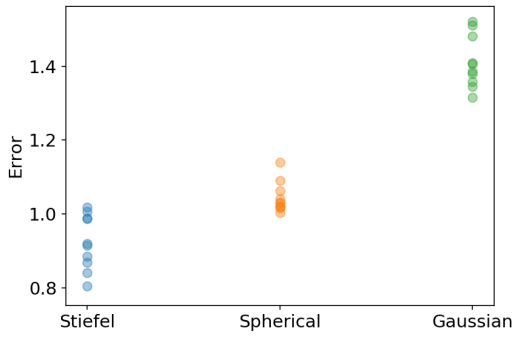
(a) $x = 0, \delta = 0.1, k = 20$



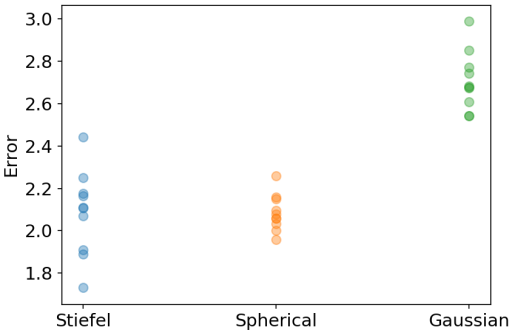
(b) $x = 0, \delta = 0.1, k = 40$



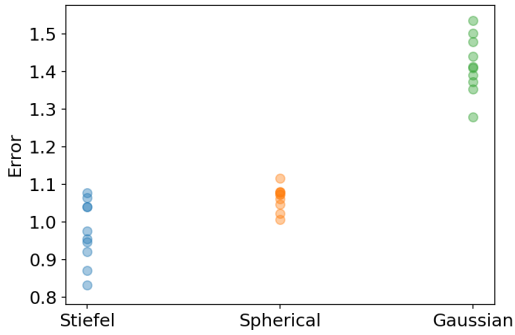
(c) $x = 0, \delta = 0.01, k = 20$



(d) $x = 0, \delta = 0.01, k = 40$

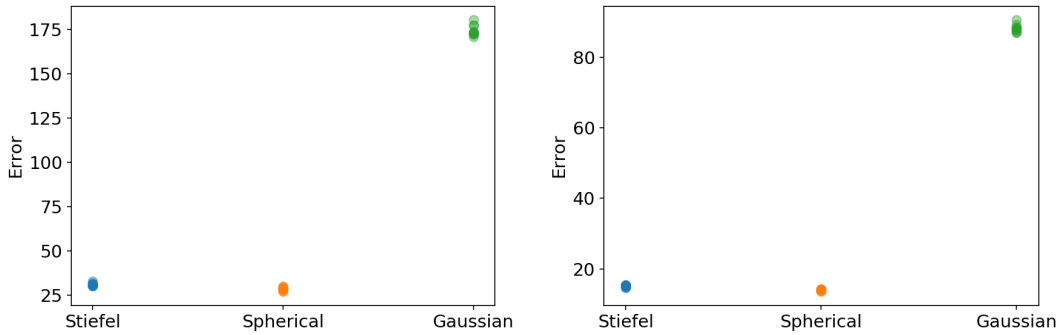


(e) $x = 0, \delta = 0.001, k = 20$

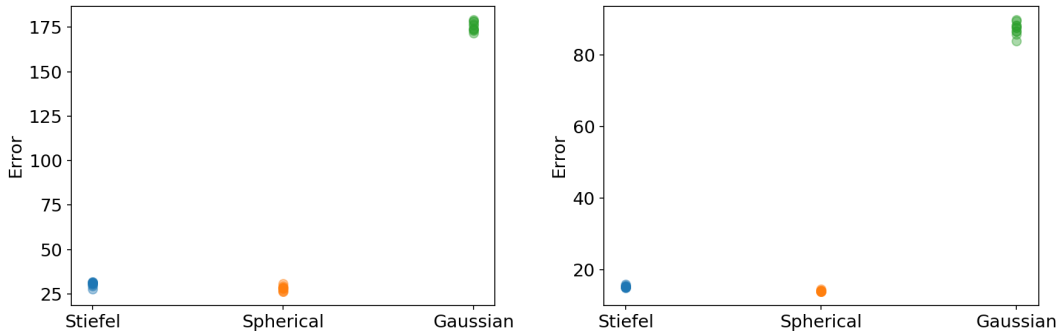


(f) $x = 0, \delta = 0.001, k = 40$

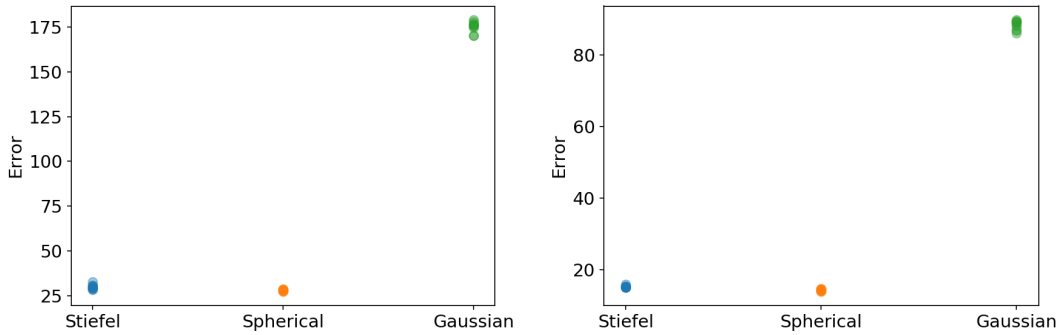
Figure 13: Errors of Hessian estimators on the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See the caption of Figure 5 for detailed illustration.



(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 20$ (b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 40$

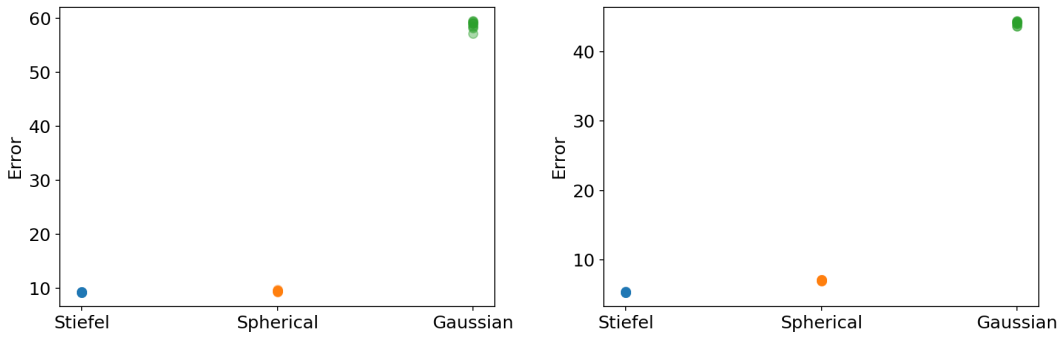


(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 20$ (d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 40$

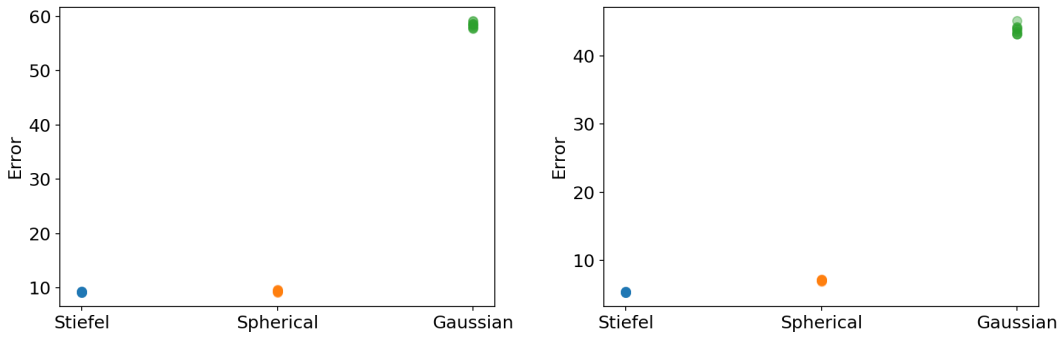


(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 20$ (f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 40$

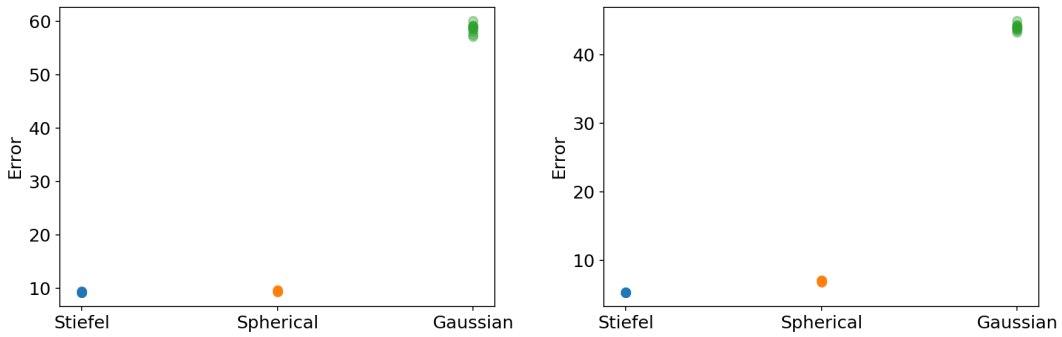
Figure 14: Errors of Hessian estimators on the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. This figure shows that when k is much smaller than n, $\widehat{\mathrm{H}}f^{\delta}_{k,S}(x)$ (Wang, 2022) can achieve same level of accuracy as $\widehat{\mathrm{H}}f^{\delta}_{k}(x)$. See the caption of Figure 5 for detailed illustration.



(a) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 60$ (b) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.1, k = 80$



(c) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 60$ (d) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.01, k = 80$



(e) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 60$ (f) $x = \frac{\pi}{4}\mathbf{1}, \delta = 0.001, k = 80$

Figure 15: Errors of Hessian estimators on the test function defined in (Eq. 7). Each subfigure corresponds to a different combination of the location for estimation x, the finite difference granularity δ, and number of random directions k. See the caption of Figure 5 for detailed illustration.