On Sharp Stochastic Zeroth Order Hessian Estimators over Riemannian Manifolds

Tianyu Wang*

Abstract

We study Hessian estimators for functions defined over an n-dimensional complete analytic Riemannian manifold. We introduce new stochastic zeroth-order Hessian estimators using $O(1)$ function evaluations. We show that, for an analytic real-valued function f, our estimator achieves a bias bound of order $O\left(\gamma\delta^2\right)$, where γ depends on both the Levi-Civita connection and function f, and δ is the finite difference step size. To the best of our knowledge, our results provide the first bias bound for Hessian estimators that explicitly depends on the geometry of the underlying Riemannian manifold. We also study downstream computations based on our Hessian estimators. The supremacy of our method is evidenced by empirical evaluations.

1 Introduction

Hessian computation is one of the central tasks in optimization, machine learning and related fields. Understanding the landscape of the objective function is in many cases the first step towards solving a mathematical programming problem, and Hessian is one of the key quantities that depict the function landscape. Often in real-world scenarios, the objective function is a black-box, and its Hessian is not directly computable. In these cases, zeroth-order Hessian computation techniques are needed if one wants to understand the function landscape via its Hessian.

To this end, we introduce new zeroth-order methods for estimating a function's Hessian at any given point over an n-dimensional complete analytic Riemannian manifold $(\mathcal{M}, g)$. For $p \in \mathcal{M}$ and an analytic real-valued function f defined over a complete analytic Riemannian manifold $\mathcal{M}$, the Hessian estimator of f at p is

$$\widehat{\mathrm{H}} f(p; v, w; \delta) := \frac{n^2}{\delta^2} f(\mathrm{Exp}_p(\delta v + \delta w)) v \otimes w, \tag{1}$$

where Exp_p is the exponential map, v, w are independently sampled from the unit sphere in $T_p\mathcal{M}$, $v \otimes w$ denotes the tensor product of v and w ($v, w \in T_p\mathcal{M}$), and δ is the finite-difference step size.

Our Hessian estimator satisfies

$$\begin{aligned} &\left\| \mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p} \left[\widehat{\mathrm{H}} f(p; v, w; \delta) \right] - \mathrm{Hess} f(p) \right\| \\ &= O\left(\delta^2 \sup_{u \in \mathbb{S}_p} \left| \mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p} \left[\frac{n}{n+2} \nabla_u^2 \left(\nabla_v^2 + \nabla_w^2 \right) f(p) \right] \right| \right), \end{aligned}$$

where $\|\cdot\|$ is the ∞-Schatten norm, $\mathbb{S}_p$ is the unit sphere in $T_p\mathcal{M}$, $\mathrm{Hess} f(p)$ is the Hessian of f at p, and ∇ is the covariant derivative associated with the Riemannian metric.

This bias bound improves previously known results in two ways:

1. It provides, via the Levi-Civita connection, the first bias bound for Hessian estimators that explicitly depends on the local geometry of the underlying space;
2. It significantly improves best previously known bias bound for $O(1)$-evaluation Hessian estimators over Riemannian manifolds, which is of order $O(L_2 n^2 \delta)$, where L_2 is the Lipschitz constant for the Hessian, n is the dimension of the manifold, and δ is the finite-difference step size. See Remark 1 for details.

*wangtianyu@fudan.edu.cn

We also study downstream computations for our Hessian estimator. More specifically, we introduce novel provably accurate methods for computing adjugate and inversion of the Hessian matrix, all using zeroth order information only. These zeroth order computation methods may be used as primers for further applications. The supremacy of our method over existing methods is evidenced by careful empirical evaluations.

Related Works

Zeroth order optimization has attracted the attention of many researchers. Under this broad umbrella, there stands the Bayesian optimization methods (See the review article by Shahriari et al. (2015) for an overview), comparison-based methods (e.g., Nelder and Mead, 1965), genetic algorithms (e.g., Goldberg and Holland, 1988), best arm identification from the multi-armed bandit community (e.g., Bubeck et al., 2009; Audibert et al., 2010), and many others (See the book by Conn et al. (2009) for an overview). Among all these zeroth order optimization schemes, one classic and prosperous line of works focuses on estimating higher order derivatives using zeroth order information.

Zeroth order gradient estimators make up a large portion of derivative estimation literature. In the past few years, Flaxman et al. (2005) studied the stochastic gradient estimator using a single-point 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); Li et al. (2020) studied gradient estimators using Gaussian smoothing, and investigated downstream optimization methods using the estimated gradient. Recently, Wang et al. (2021) studied stochastic gradient estimators over Riemannian manifolds, via the lens of the Greene-Wu convolution.

Zeroth order Hessian estimation is also a central topic in derivative estimation. In the control community, gradient-based Hessian estimators were introduced for iterative optimization algorithms, and asymptotic convergence was proved (Spall, 2000). Apart from this asymptotic result, no generic non-asymptotic bound for $O(1)$-evaluation Hessian estimators are well investigated until recently. Based on the Stein's identity (Stein, 1981), Balasubramanian and Ghadimi (2021) designed the Stein-type Hessian estimator, and combined it with cubic regularized Newton's method (Nesterov and Polyak, 2006) for non-convex optimization. Li et al. (2020) generalizes the Stein-type Hessian estimators to Riemannian manifolds embedded in Euclidean spaces. Several authors have also considered variance and higher order moments of Hessian (and gradient) estimators (Li et al., 2020; Balasubramanian and Ghadimi, 2021; Feng and Wang, 2022). In particular, Feng and Wang (2022) showed that estimators via random orthogonal frames from Steifel's manifolds have significantly smaller variance. Yet in the case of non-trivial curvature (Li et al., 2020), no geometry-aware bias bound has been given prior to our work.

2 Preliminaries and Conventions

For better readability, we list here some notations and conventions that will be used throughout the rest of this paper.

- For any $p \in \mathcal{M}$, let U_p denote the open set near p that is diffeomorphic to a subset of $\mathbb{R}^n$ via the local normal coordinate chart ϕ. Define the distance $d_p(q_1, q_2)$ $(q_1, q_2 \in U_p)$ such that

$$d_p(q_1, q_2) = d_{\text{Euc}}(\phi(q_1), \phi(q_2)).$$

  where d_{Euc} is the Euclidean distance in $\mathbb{R}^n$.

- **(A0, Analyticity Assumption)** Throughout the paper, we assume that, both the Riemannian metric and the function of interest are analytic.

- The injectivity radius of $p \in \mathcal{M}$ (written $\text{inj}(p)$) is defined as the radius of the largest geodesic ball that is contained in U_p. **(A1, Small Step Size Assumption)** Throughout the paper, we assume that the finite difference step size δ of the estimator at point $p \in \mathcal{M}$ satisfies $\delta \leq \frac{\text{inj}(p)}{2}$.

- All musical isomorphisms are omitted when there is no confusion.
- For any $p \in \mathcal{M}$ and $\alpha > 0$, we use $\alpha\mathbb{S}_p$ (resp. $\alpha\mathbb{B}_p$) to denote the origin-centered sphere (resp. ball) in $T_p\mathcal{M}$ with radius α. For simplicity, we write $\mathbb{S}_p = 1\mathbb{S}_p$ (resp. $\mathbb{B}_p = 1\mathbb{B}_p$). It is worth emphasizing that $\mathbb{S}_p$ and $\mathbb{B}_p$ are in $T_p\mathcal{M}$. They are different from geodesic balls which reside in $\mathcal{M}$.
- For $p \in \mathcal{M}$ and $q \in U_p$, we use $\mathcal{I}_p^q : T_p\mathcal{M} \to T_q\mathcal{M}$ to denote the parallel transport from $T_p\mathcal{M}$ to $T_q\mathcal{M}$ along the distance-minimizing geodesic connecting p and q. For any $p \in \mathcal{M}$, $u \in T_p\mathcal{M}$ and $q \in U_p$, define $u_q = \mathcal{I}_p^q(u)$. More generally, $\mathcal{I}_p^q$ denotes the parallel transport along the distance-minimizing geodesic from p to q, among the fiber bundle that is compatible with the Riemannian structure.
- We will use the double exponential map notation (Gavrilov, 2007). For any $p \in \mathcal{M}$ and $u, v \in T_p\mathcal{M}$ such that $\mathrm{Exp}_p(u) \in U_p$, we write $\mathrm{Exp}_p^2(u, v) = \mathrm{Exp}_{\mathrm{Exp}_p(u)}(v_{\mathrm{Exp}_p(u)})$.
- **(Definition of Hessian (e.g., Petersen, 2006))** Over an n-dimensional complete Riemannian manifold $\mathcal{M}$, the Hessian of a smooth function $f : \mathcal{M} \to \mathbb{R}$ at p is a bilinear form $\mathrm{Hess} f(p) : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ such that, for all $u, v \in T_p\mathcal{M}$, $\mathrm{Hess} f(p)(u, v) = \left\langle \nabla_v df\big|_p, u \right\rangle$. Since the Levi-Civita connection is torsion-free, the Hessian is symmetric: $\mathrm{Hess} f(p)(u, v) = \mathrm{Hess} f(p)(v, u)$ for all $u, v \in T_p\mathcal{M}$. For a smooth function f, its Hessian satisfies (e.g., Chapter 5.4 of (Absil et al., 2009)), for any $p \in \mathcal{M}$ and any $v \in T_p\mathcal{M}$,

$$\mathrm{Hess} f(p)(v, v) = \lim_{\tau \to 0} \frac{f(\mathrm{Exp}_p(\tau v)) - 2f(p) + f(\mathrm{Exp}_p(-\tau v))}{\tau^2} = \nabla_v^2 f(p). \tag{2}$$

  For simplicity and coherence with the notations in the Euclidean case, we write $u^\top \mathrm{Hess} f(p) v := \mathrm{Hess} f(p)(u, v)$ for any $u, v \in T_p\mathcal{M}$.
- Consider a Riemannian manifold $(\mathcal{M}, g)$, a point $p \in \mathcal{M}$, and any symmetric bilinear form $A : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$. The g-induced ∞-Schatten norm (the operator norm) of A is defined as

$$\|A\| = \sup_{u \in T_p\mathcal{M}, \|u\|=1} |u^\top A u|.$$

  When it is clear from context, we simply use ∞-Schatten norm to refer to g-induced ∞-Schatten norm.
- *Note.* When applied to a tangent vector, $\|\cdot\|$ is the standard norm induced by the Riemannian metric. When applied to a symmetric bilinear form, $\|\cdot\|$ is the ∞-Schatten norm defined above.

3 Zeroth Order Hessian Estimation

For $p \in \mathcal{M}$ and $f : \mathcal{M} \to \mathbb{R}$, the Hessian of f at p can be estimated by

$$\widehat{\mathrm{H}} f(p; v, w; \delta) = \frac{n^2}{\delta^2} f(\mathrm{Exp}_p(\delta v + \delta w)) v \otimes w,$$

where v, w are independently uniformly sampled from $\mathbb{S}_p$ and δ is the finite difference step size. To study the bias of this estimator, we consider a function $\widetilde{f}^\delta$ defined as follows.

For $p \in \mathcal{M}$, a smooth real-valued function f defined over $\mathcal{M}$, and a number $\delta \in (0, \delta_0]$, define a function $\widetilde{f}^\delta$ (at p) such that

$$\widetilde{f}^\delta(p) = \frac{1}{\delta^{2n} V_n^2} \int_{w \in \delta\mathbb{B}_p} \int_{v \in \delta\mathbb{B}_p} f(\mathrm{Exp}_p(v + w))\, dw\, dv, \tag{3}$$

where V_n is the volume of the unit ball in $T_p\mathcal{M}$ (same as the volume of the unit ball in $\mathbb{R}^n$). Smoothings of this kind have been analytically investigated by Greene and Wu (Greene and Wu, 1973, 1976, 1979). We will first show that $\mathrm{Hess}\widetilde{f}^{\delta}(p) = \mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p}\left[\widehat{\mathrm{H}}f(p;v,w;\delta)\right]$ in Lemma 1. Then we derive a bound on $\left\|\mathrm{Hess}\widetilde{f}^{\delta}(p) - \mathrm{Hess}f(p)\right\|$, which gives a bound on $\left\|\mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p}\left[\widehat{\mathrm{H}}f(p;v,w;\delta)\right] - \mathrm{Hess}f(p)\right\|$. Henceforth, we will use $\mathbb{E}_{v,w}$ as a shorthand for $\mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p}$.

Lemma 1. *Consider an n-dimensional complete analytic Riemannian manifold $(\mathcal{M}, g)$. Consider $p \in \mathcal{M}$, an analytic function $f : \mathcal{M} \to \mathbb{R}$ and a small number $\delta \in (0, \mathrm{inj}(p)/2]$. If v and w are independently randomly sampled from $\mathbb{S}_p$, then it holds that,*

$$\mathbb{E}_{v,w}\left[\widehat{\mathrm{H}}f(p;v,w;\delta)\right] = \mathrm{Hess}\widetilde{f}^{\delta}(p).$$

Proof. Define $\varphi_p = f \circ \mathrm{Exp}_p$. By the fundamental theorem of geometric calculus, it holds that [1]

$$\begin{aligned}\int_{v\in\delta\mathbb{B}_p} \partial_i \int_{w\in\delta\mathbb{B}_p} \partial_j \varphi_p(w+v)\, dw\, dv &= \int_{v\in\delta\mathbb{B}_p} \partial_i \int_{w\in\delta\mathbb{S}_p} \varphi_p(w+v)\frac{w}{\|w\|}\, dw\, dv \\ &\overset{(i)}{=} \int_{v\in\delta\mathbb{S}_p}\int_{w\in\delta\mathbb{S}_p} \varphi_p(w+v)\frac{v\otimes w}{\|w\|\|v\|}\, dw\, dv.\end{aligned}$$

Since v and w are independently uniformly sampled from $\mathbb{S}_p$, it holds that

$$\int_{\delta\mathbb{S}_p}\int_{\delta\mathbb{S}_p} \varphi_p(w+v)\frac{v\otimes w}{\|v\|\|w\|}\, dw\, dv \overset{(ii)}{=} \delta^{2n-2}A_{n-1}^2\mathbb{E}_{v,w}\left[\varphi_p(\delta v+\delta w)v\otimes w\right],$$

where A_{n-1} is the surface area of $\mathbb{S}_p$ in $T_p\mathcal{M}$ (same as the surface area of unit sphere in $\mathbb{R}^n$).

By the dominated convergence theorem and that $\delta \le \frac{\mathrm{inj}(p)}{2}$, we have

$$\partial_i\partial_j \int_{v\in\delta\mathbb{B}_p}\int_{w\in\delta\mathbb{B}_p} \varphi_p(w+v)\, dw\, dv \overset{(iii)}{=} \int_{v\in\delta\mathbb{B}_p} \partial_i \int_{w\in\delta\mathbb{B}_p} \partial_j\varphi_p(w+v)\, dw\, dv.$$

More specifically, the ∂_i operations (or tangent vectors) can be defined in terms of limits, and we can interchange the limit and the integral by the dominated convergence theorem.

Combining (i), (ii) and (iii) gives

$$\partial_i\partial_j \int_{v\in\delta\mathbb{B}_p}\int_{w\in\delta\mathbb{B}_p} \varphi_p(w+v)\, dw\, dv \overset{(iv)}{=} \delta^{2n-2}A_{n-1}^2\mathbb{E}_{v,w}\left[\varphi_p(\delta v+\delta w)v\otimes w\right].$$

Combining the above results gives

$$\begin{aligned}\partial_i\partial_j\widetilde{f}^{\delta}(p) &= \partial_i\partial_j \frac{1}{\delta^{2n}V_n}\int_{v\in\delta\mathbb{B}_p}\int_{w\in\delta\mathbb{B}_p} \varphi_p(w+v)\, dw\, dv \\ &= \frac{\delta^{2n-2}A_{n-1}^2}{\delta^{2n}V_n^2}\mathbb{E}_{v,w}\left[\varphi_p(\delta v+\delta w)v\otimes w\right] \\ &= \frac{n^2}{\delta^2}\mathbb{E}_{v,w}\left[f(\mathrm{Exp}_p(\delta v+\delta w))v\otimes w\right],\end{aligned}$$

where the second last equality uses (iv), and last equality uses $A_{n-1} = nV_n$. □

As a result of Lemma 1, a bound on $\|\mathrm{Hess}\widetilde{f}^{\delta}(p) - \mathrm{Hess}f(p)\|$ will give a bound on $\|\mathbb{E}\left[\widehat{\mathrm{H}}f(p;v,w;\delta)\right] - \mathrm{Hess}f(p)\|$. To bound $\|\mathrm{Hess}\widetilde{f}^{\delta}(p) - \mathrm{Hess}f(p)\|$, we need to explicitly extend the definition of $\widetilde{f}^{\delta}$ from p to a neighborhood of p (Wang et al., 2021), so that the Hessian can be computed in a precise manner. For $p \in \mathcal{M}$, a smooth function $f : \mathcal{M} \to \mathbb{R}$, and a number $\delta \in (0, \delta_0]$, define a function $\widetilde{f}^{\delta}$ (near p) such that

$$\widetilde{f}^{\delta}(q) = \mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p}\left[\widetilde{f}^{\delta}_{v,w}(q)\right], \quad \forall q \in U_p, \tag{4}$$

[1] Here ∂_i and ∂_j are understood as Einstein's notations.

where

$$\widetilde{f}^{\delta}_{v,w}(q) := \frac{n^2}{4\delta^{2n}} \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} f\left(\mathrm{Exp}_q(tv_q + sw_q)\right) |t|^{n-1}|s|^{n-1}\, dt\, ds, \tag{5}$$

with $v, w \in \mathbb{S}_p$.

The advantage of defining $\widetilde{f}^{\delta}$ via $\widetilde{f}^{\delta}_{v,w}$ is that $\widetilde{f}^{\delta}_{v,w}$ is explicitly defined in a neighborhood of p. Thus we can carry out geometry-aware computations in a precise manner. Next, we verify that (3) and (4) agree with each other in the following proposition.

Proposition 1. *For any $p \in \mathcal{M}$ and any $\delta \le \delta_0$, (3) and (4) coincide at any $q \in U_p$.*

Proof. At any $q \in U_p$, we have

$$\begin{aligned}
(3) &= \frac{1}{\delta^{2n} V_n^2} \int_{w\in\delta\mathbb{B}_q}\int_{v\in\delta\mathbb{B}_q} f(\mathrm{Exp}_q(v+w))\, dv\, dw \\
&\overset{(i)}{=} \frac{n^2}{\delta^{2n} A_n^2} \int_{w\in\delta\mathbb{B}_q}\int_{v\in\delta\mathbb{B}_q} f(\mathrm{Exp}_q(v+w))\, dv\, dw \\
&\overset{(ii)}{=} \frac{n^2}{\delta^{2n} A_n^2} \int_{w\in\mathbb{S}_q}\int_{v\in\mathbb{S}_q} \frac{1}{4}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} f(\mathrm{Exp}_q(tv+sw))|t|^{n-1}|s|^{n-1}\, dt\, ds\, dv\, dw,
\end{aligned}$$

where (i) uses $A_{n-1} = nV_n$, and (ii) changes from Cartesian coordinate to hyperspherical coordinate (in $T_q\mathcal{M}$). Since the Levi-Civita connection is compatible with the Riemannian metric, we know that the standard Lebesgue measure in $T_p\mathcal{M}$ is preserved after transporting to $T_q\mathcal{M}$. This implies, for any continuous function h defined over $T_q\mathcal{M}$, we have $\int_{v\in\mathbb{S}_q} h(v)dv \overset{(iii)}{=} \int_{v\in\mathbb{S}_p} h(v_q)dv$. Thus we have, at any $q \in \mathcal{M}$,

$$\begin{aligned}
(3) &= \frac{n^2}{4\delta^{2n} A_n^2} \int_{w\in\mathbb{S}_q}\int_{v\in\mathbb{S}_q}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} f(\mathrm{Exp}_q(tv+sw))|t|^{n-1}|s|^{n-1}\, dt\, ds\, dw\, dv \\
&\overset{(iv)}{=} \frac{n^2}{4\delta^{2n} A_n^2} \int_{w\in\mathbb{S}_p}\int_{v\in\mathbb{S}_p}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} f(\mathrm{Exp}_q(tv_q+sw_q))|t|^{n-1}|s|^{n-1}\, dt\, ds\, dw\, dv \\
&= \frac{1}{A_n^2}\int_{w\in\mathbb{S}_p}\int_{v\in\mathbb{S}_p} \widetilde{f}^{\delta}_{v,w}(q)\, dw\, dv = (4),
\end{aligned}$$

where (iv) uses (iii). □

By Proposition 1, it is sufficient to work with $\widetilde{f}^{\delta}_{v,w}$ and randomize v, w over a unit sphere. For any direction $u \in \mathbb{S}_p$, the Hessian of $\widetilde{f}^{\delta}_{v,w}$ along u can be explicitly written out in terms of f and u, v, w. This result is found in Lemma 2.

Lemma 2. *Consider an n-dimensional complete analytic Riemannian manifold $(\mathcal{M}, g)$. Consider $p \in \mathcal{M}$, an analytic function $f : \mathcal{M} \to \mathbb{R}$ and a small number $\delta \in (0, \mathrm{inj}(p)/2]$. For any $u, v, w \in \mathbb{S}_p$, we have*

$$\begin{aligned}
&u^\top \mathrm{Hess}\widetilde{f}^{\delta}_{v,w}(p)u \\
&= \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} u_q^\top \mathrm{Hess} f(q) u_q\ |t|^{n-1}|s|^{n-1}\, dt\, ds \\
&\quad + \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\nabla_u^2\left(t\nabla_v + s\nabla_w\right)^{2j} f(p)\, dt\, ds \\
&\quad - \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\left(t\nabla_v + s\nabla_w\right)^{2j}\nabla_u^2 f(p)\, dt\, ds,
\end{aligned}$$

where $q = \mathrm{Exp}_p(tv + sw)$.

Proof. From the definition of Hessian, we have

$$\begin{aligned} & u^\top \mathrm{Hess} \widetilde{f}^{\delta}_{v,w}(p) u \\ = & \lim_{\tau \to 0} \frac{\widetilde{f}^{\delta}_{v,w}(\mathrm{Exp}_p(\tau u)) - 2\widetilde{f}^{\delta}_{v,w}(p) + \widetilde{f}^{\delta}_{v,w}(\mathrm{Exp}_p(-\tau u))}{\tau^2}. \end{aligned}$$

Thus it is sufficient to fix any $t, s \in [-\delta, \delta]$ and consider

$$\lim_{\tau \to 0} \frac{f(\mathrm{Exp}_p^2(\tau u, tv + sw)) - 2f(\mathrm{Exp}_p(tv + sw)) + f(\mathrm{Exp}_p^2(-\tau u, tv + sw))}{\tau^2}.$$

For simplicity, define

$$\begin{aligned} \phi(\tau, t, s) = & f(\mathrm{Exp}_p^2(\tau u, tv + sw)) + f(\mathrm{Exp}_p^2(-\tau u, tv + sw)) \\ & - f(\mathrm{Exp}_p^2(tv + sw, \tau u)) - f(\mathrm{Exp}_p^2(tv + sw, -\tau u)). \end{aligned}$$

Let $q = \mathrm{Exp}_p(tv + sw)$, and we have

$$\begin{aligned} & u^\top \mathrm{Hess} \widetilde{f}^{\delta}_{v,w}(p) u \\ = & \frac{n^2}{4\delta^{2n}} \int_{-\delta}^{\delta} \int_{-\delta}^{\delta} u_q^\top \mathrm{Hess} f(q) u_q \, |t|^{n-1} |s|^{n-1} \, dt \, ds \\ & + \frac{n^2}{4\delta^{2n}} \lim_{\tau \to 0} \frac{\int_{-\delta}^{\delta} \int_{-\delta}^{\delta} \phi(\tau, t, s) |t|^{n-1} |s|^{n-1} \, dt \, ds}{\tau^2}, \end{aligned} \tag{6}$$

provided that the last term converges.

For any $p \in \mathcal{M}$, $v \in T_p\mathcal{M}$ and $q \in U_p$, define $h_v^{(j)}(q) = \nabla_{v_q}^j f(q)$. We can Taylor expand $h_v^{(j)}(\mathrm{Exp}_p(u))$ by

$$\begin{aligned} h_v^{(j)}(\mathrm{Exp}_p(u)) &= h_v^{(j)}(\mathrm{Exp}_p(tu))\big|_{t=1} \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} \frac{d^i}{dt^i} h_v^{(j)}(\mathrm{Exp}_p(tu))\Big|_{t=0} \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} \nabla_u^i h_v^{(j)}(p) \\ &\overset{(a)}{=} \sum_{i=0}^{\infty} \frac{1}{i!} \nabla_u^i \nabla_v^j f(p). \end{aligned}$$

From above, we have, for any p, and $u, v \in T_p\mathcal{M}$ of small norm,

$$\begin{aligned} f\left(\mathrm{Exp}_p^2(u, v)\right) &= f\left(\mathrm{Exp}_{\mathrm{Exp}_p(u)}\left(v_{\mathrm{Exp}_p(u)}\right)\right) \\ &= \sum_{j=0}^{\infty} \frac{1}{j!} \nabla^j_{v_{\mathrm{Exp}_p(u)}} f(\mathrm{Exp}_p(u)) \\ &= \sum_{j=0}^{\infty} \frac{1}{j!} h_v^{(j)}(\mathrm{Exp}_p(u)) \\ &= \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{i=0}^{\infty} \frac{1}{i!} \nabla_u^i \nabla_v^j f(p), \end{aligned}$$

where the second equality uses Taylor expansion at $\mathrm{Exp}_p(u)$ and the last equality uses (a).

From the above computation, we expand $f(\mathrm{Exp}_p^2(tv+sw,\tau u))$ (and similar terms) into infinite series. Thus we can write $\phi(\tau,t,s)$ as

$$\begin{aligned}
\phi(\tau,t,s) &= f(\mathrm{Exp}_p^2(\tau u, tv+sw)) + f(\mathrm{Exp}_p^2(-\tau u, tv+sw)) \\
&\quad - f(\mathrm{Exp}_p^2(tv+sw,\tau u)) - f(\mathrm{Exp}_p^2(tv+sw,-\tau u)) \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{i!j!}\nabla_{\tau u}^i\nabla_{tv+sw}^j f(p) + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{i!j!}\nabla_{-\tau u}^i\nabla_{tv+sw}^j f(p) \\
&\quad - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{i!j!}\nabla_{tv+sw}^j\nabla_{\tau u}^i f(p) - \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{1}{i!j!}\nabla_{tv+sw}^j\nabla_{-\tau u}^i f(p) \\
&= \sum_{j=0}^{\infty}\frac{\tau^2}{2j!}\nabla_u^2\nabla_{tv+sw}^j f(p) + \sum_{j=0}^{\infty}\frac{\tau^2}{2j!}\nabla_u^2\nabla_{tv+sw}^j f(p) \\
&\quad - \sum_{j=0}^{\infty}\frac{\tau^2}{2j!}\nabla_{tv+sw}^j\nabla_u^2 f(p) - \sum_{j=0}^{\infty}\frac{\tau^2}{2j!}\nabla_{tv+sw}^j\nabla_u^2 f(p) + O(\tau^3),
\end{aligned} \tag{7}$$

where the last equality uses that zeroth-order terms in τ and first-order terms in τ all cancel.

From (7), we have

$$\begin{aligned}
\lim_{\tau\to 0}\frac{\phi(\tau,t,s)}{\tau^2} &= \sum_{j=1}^{\infty}\frac{1}{j!}\nabla_u^2\nabla_{tv+sw}^j f(p) - \sum_{j=1}^{\infty}\frac{1}{j!}\nabla_{tv+sw}^j\nabla_u^2 f(p) \\
&= \sum_{j=1}^{\infty}\frac{1}{j!}\nabla_u^2\left(t\nabla_v + s\nabla_w\right)^j f(p) - \sum_{j=1}^{\infty}\frac{1}{j!}\left(t\nabla_v + s\nabla_w\right)^j\nabla_u^2 f(p) \\
&= \sum_{j=1}^{\infty}\frac{1}{(2j)!}\nabla_u^2\left(t\nabla_v + s\nabla_w\right)^{2j} f(p) \\
&\quad - \sum_{j=1}^{\infty}\frac{1}{(2j)!}\left(t\nabla_v + s\nabla_w\right)^{2j}\nabla_u^2 f(p) + Odd(t,s),
\end{aligned}$$

where $Odd(t,s)$ denotes terms that are either odd in t or odd in s.

Since $\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} Odd(t,s)\,dt\,ds = 0$, we have

$$\begin{aligned}
&\frac{n^2}{4\delta^{2n}}\lim_{\tau\to 0}\frac{\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\phi(\tau,t,s)|t|^{n-1}|s|^{n-1}\,dt\,ds}{\tau^2} \\
&= \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty}\frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\nabla_u^2\left(t\nabla_v + s\nabla_w\right)^{2j} f(p)\,dt\,ds \\
&\quad - \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty}\frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\left(t\nabla_v + s\nabla_w\right)^{2j}\nabla_u^2 f(p)\,dt\,ds.
\end{aligned} \tag{8}$$

Collecting terms from (6) and (8), we have

$$\begin{aligned}
&u^\top \mathrm{Hess}\widetilde{f}_{v,w}^{\delta}(p)u \\
&= \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} u_q^\top \mathrm{Hess} f(q) u_q\ |t|^{n-1}|s|^{n-1}\,dt\,ds \\
&\quad + \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty}\frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\nabla_u^2\left(t\nabla_v + s\nabla_w\right)^{2j} f(p)\,dt\,ds \\
&\quad - \frac{n^2}{4\delta^{2n}}\int_{-\delta}^{\delta}\int_{-\delta}^{\delta}\sum_{j=1}^{\infty}\frac{|t|^{n-1}|s|^{n-1}}{(2j)!}\left(t\nabla_v + s\nabla_w\right)^{2j}\nabla_u^2 f(p)\,dt\,ds,
\end{aligned}$$

where $q = \mathrm{Exp}_p(tv+sw)$. This concludes the proof. □

Gathering the above results gives a bias bound for (1), which is summarized in the following theorem.

Theorem 1. *Consider an n-dimensional complete analytic Riemannian manifold $(\mathcal{M}, g)$. Consider $p \in \mathcal{M}$, an analytic function $f : \mathcal{M} \to \mathbb{R}$ and a small number $\delta \in (0, \mathrm{inj}(p)/2]$. For any $p \in \mathcal{M}$ and unit vectors $u, v \in T_p\mathcal{M}$, define a function $\vartheta_{p,u,v,w}$ over $(-\mathrm{inj}(p)/2, \mathrm{inj}(p)/2) \times (-\mathrm{inj}(p)/2, \mathrm{inj}(p)/2)$ such that*

$$\vartheta_{p,u,v,w}(t,s) = \mathrm{Hess} f(\mathrm{Exp}_p(tv+sw))(u_{\mathrm{Exp}_p(tv+sw)}, u_{\mathrm{Exp}_p(tv+sw)}).$$

The estimator (1) satisfies

$$\begin{aligned}
&\left\| \mathbb{E}_{v,w}\left[\widehat{\mathrm{H}} f(p; v, w; \delta)\right] - \mathrm{Hess} f(p) \right\| \\
\leq &\sup_{u \in \mathbb{S}_p} \left| \mathbb{E}_{v,w}\left[\frac{n^2}{4\delta^{2n}} \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{i,j\in\mathbb{N}, i+j\geq 1} \frac{t^{2i}s^{2j}}{(2i)!(2j)!} \partial_1^{2i}\partial_2^{2j}\vartheta_{p,u,v,w}(0,0)|t|^{n-1}|s|^{n-1}\, dt\, ds \right] \right. \\
&+ \mathbb{E}_{v,w}\left[\frac{n^2}{4\delta^{2n}} \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!} \nabla_u^2 \left(t\nabla_v + s\nabla_w\right)^{2j} f(p)\, dt\, ds \right. \\
&\left.\left. - \frac{n^2}{4\delta^{2n}} \int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!} \left(t\nabla_v + s\nabla_w\right)^{2j} \nabla_u^2 f(p)\, dt\, ds \right] \right|,
\end{aligned}$$

where v, w are independently sampled from the uniform distribution over $\mathbb{S}_p$.

Proof. By Lemma 2, we have

$$\begin{aligned}
&u^\top \mathbb{E}_{v,w}\left[\mathrm{Hess}\widetilde{f}^{\delta}_{v,w}(p)\right] u \\
= &\frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} u^\top_{\mathrm{Exp}_p(tv+sw)} \mathrm{Hess} f(\mathrm{Exp}_p(tv+sw)) u_{\mathrm{Exp}_p(tv+sw)}\ |t|^{n-1}|s|^{n-1}\, dt\, ds \right] \\
&+ \frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!} \nabla_u^2 \left(t\nabla_v + s\nabla_w\right)^{2j} f(p)\, dt\, ds \right] \\
&- \frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{j=1}^{\infty} \frac{|t|^{n-1}|s|^{n-1}}{(2j)!} \left(t\nabla_v + s\nabla_w\right)^{2j} \nabla_u^2 f(p)\, dt\, ds \right].
\end{aligned}$$

By the analyticity assumption, we have

$$\vartheta_{p,u,v,w}(t,s) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{t^i s^j}{i!j!} \partial_1^i \partial_2^j \vartheta_{p,u,v,w}(0,0).$$

For any fixed $u \in \mathbb{S}_p$, we have

$$\begin{aligned}
&\frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} u^\top_{\mathrm{Exp}_p(tv+sw)} \mathrm{Hess} f(\mathrm{Exp}_p(tv+sw)) u_{\mathrm{Exp}_p(tv+sw)}\ |t|^{n-1}|s|^{n-1}\, dt\, ds \right] \\
= &\frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \vartheta_{p,u,v,w}(t,s)\ |t|^{n-1}|s|^{n-1}\, dt\, ds \right] \\
= &\frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{t^i s^j}{i!j!} \partial_1^i \partial_2^j \vartheta_{p,u,v,w}(0,0) \right) |t|^{n-1}|s|^{n-1}\, dt\, ds \right] \\
= &\frac{n^2}{4\delta^{2n}} \mathbb{E}_{v,w}\left[\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} \sum_{i,j\in\mathbb{N}, i+j\geq 1} \frac{t^{2i}s^{2j}}{(2i)!(2j)!} \partial_1^{2i}\partial_2^{2j}\vartheta_{p,u,v,w}(0,0)|t|^{n-1}|s|^{n-1}\, dt\, ds \right] + u^\top \mathrm{Hess} f(p) u,
\end{aligned}$$

where in the last line the terms that are odd in t or s vanishes. □

By applying (2) twice, we have

$$\partial_1^2 \vartheta_{p,u,v,w}(0,0) = \nabla_v^2 \nabla_u^2 f(p) \qquad \text{and} \qquad \partial_2^2 \vartheta_{p,u,v,w}(0,0) = \nabla_w^2 \nabla_u^2 f(p).$$

Thus by dropping terms of order $O(\delta^3)$ and noting that $\int_{-\delta}^{\delta}\int_{-\delta}^{\delta} |t|^{n-1}t|s|^{n-1}s \; dt \; ds = 0$, we have

$$\begin{aligned}
&\left\| \mathbb{E}_{v,w}\left[\widehat{\mathrm{H}} f(p; v, w; \delta)\right] - \mathrm{Hess} f(p) \right\| \\
&= O\left(\delta^2 \sup_{u \in \mathbb{S}_p} \left| \mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}_p} \left[\frac{n}{n+2} \nabla_u^2 \left(\nabla_v^2 + \nabla_w^2 \right) f(p) \right] \right| \right).
\end{aligned}$$

3.1 Example: the n-sphere

We consider the Riemannian manifold $\mathbb{S}^{n-1}$ with metric induced by the ambient Euclidean space. This space of both theoretical and practical appeal. In this space, the exponential map is

$$\mathrm{Exp}_x(tv) = x\cos(t) + v\sin(t),$$

where v is a unit vector in $\mathbb{R}^n$; The parallel transport is

$$\mathcal{I}_x^{\mathrm{Exp}_x(tu)}(v) = v - uu^\top v + uu^\top v \cos(t) - \|uu^\top v\| x \sin(t)$$

for any $x \in \mathbb{S}^{n-1}$, $u, v \in \mathbb{S}^{n-1}$ and $u, v \perp x$.

We will consider estimating Hessian of the function x_i^2 where $x \in \mathbb{S}^{n-1}$ and x_i is the i-th component of x. This simple function serves as an example of estimating the Hessian of general polynomials over $\mathbb{S}^{n-1}$.

We have

$$\begin{aligned}
\nabla_v^2 x_i^2 &= \lim_{t\to 0} \frac{\left(x_i \cos t + v_i \sin t\right)^2 - 2x_i^2 + \left(x_i \cos t - v_i \sin t\right)^2}{t^2} \\
&= \lim_{t\to 0} \frac{\left(2\cos^2 t - 2\right) x_i^2 + 2v_i^2 \sin^2 t}{t^2} \\
&= -2x_i^2 + 2v_i^2,
\end{aligned}$$

and

$$\begin{aligned}
&\nabla_u^2 v_i^2 \\
&= \lim_{t\to 0} \frac{2\left(v_i - \left(uu^\top v\right)_i + \left(uu^\top v\right)_i \cos t\right)^2 + 2\left(\|uu^\top v\| x_i \sin t\right)^2 - 2v_i^2}{t^2} \\
&= \lim_{t\to 0} \frac{4v_i \left(uu^\top v\right)_i (\cos t - 1) + 2(uu^\top v)_i^2 (\cos t - 1)^2 + 2\left(\|uu^\top v\| x_i \sin t\right)^2}{t^2} \\
&= -2v_i \left(uu^\top v\right)_i + 2\|uu^\top v\|^2 x_i^2.
\end{aligned}$$

Thus it holds that

$$\begin{aligned}
\nabla_u^2 \nabla_v^2 x_i^2 &= \nabla_u^2 \left(-2x_i^2 + 2v_i^2\right) \\
&= -2\left(-2x_i^2 + 2u_i^2\right) + 2\left(-2v_i \left(uu^\top v\right)_i + 2\|uu^\top v\|^2 x_i^2\right) \\
&= 4x_i^2 - 4u_i^2 - 4v_i \left(uu^\top v\right)_i + 4\|uu^\top v\|^2 x_i^2.
\end{aligned}$$

Since $\mathbb{E}_v\left[vv^\top\right] = \frac{1}{n} I$ and $\mathbb{E}_w\left[ww^\top\right] = \frac{1}{n} I$, we have

$$\begin{aligned}
\mathbb{E}_{v,w}\left[\nabla_u^2 \left(\nabla_v^2 + \nabla_w^2\right) x_i^2\right] &= 2\left(4x_i^2 - 4u_i^2 - \frac{4}{n}u_i^2 + \frac{4}{n}x_i^2\right) \\
&= \left(8 + \frac{8}{n}\right) x_i^2 - \left(8 + \frac{8}{n}\right) u_i^2.
\end{aligned}$$

This implies that the Hessian estimator for x_i^2 over the n-sphere with granularity δ is of order

$$O\left(\delta^2 \max_{u \in \mathbb{S}^{n-1}, u \perp x} \left| \left(8 + \frac{8}{n}\right) x_i^2 - \left(8 + \frac{8}{n}\right) u_i^2 \right| \right).$$

4 The Euclidean Case

In this section, we will focus on numerical stabilization of the estimation, and algorithmic zeroth-order inversion of the estimated Hessian. For numerical and algorithmic purposes, we restrict our attention to the Euclidean case. In the Euclidean case, we also use the notation $\nabla^2 f(x)$ to denote the Hessian of f at x.

4.1 Stabilizing the Estimate

In the Euclidean case, the estimator in (1) simplifies to

$$\widehat{\mathrm{H}} f(p; v, w; \delta) = \frac{n^2}{\delta^2} f(p + \delta v + \delta w) v w^\top,$$

where v, w are independently uniformly sampled from $\mathbb{S}^{n-1}$ (the unit sphere in $\mathbb{R}^n$). Its stabilized version is

$$\begin{aligned}\widehat{\mathrm{H}} f(p; v, w; \delta) = \frac{n^2}{8\delta^2} \big[& f(p + \delta v + \delta w) - f(p - \delta v + \delta w) \\ & - f(p + \delta v - \delta w) + f(p - \delta v - \delta w) \big] \left(v w^\top + w v^\top \right). \end{aligned} \tag{9}$$

To see why (9) stabilizes the estimate, we use Taylor expansion and get

$$\begin{aligned} & f(p + \delta v + \delta w) - f(p - \delta v + \delta w) - f(p + \delta v - \delta w) + f(p - \delta v - \delta w) \\ \approx\ & \delta^2 \left(v + w \right)^\top \nabla^2 f(x) \left(v + w \right) - \frac{\delta^2}{2} \left(v - w \right)^\top \nabla^2 f(x) \left(v - w \right) \\ & - \frac{\delta^2}{2} \left(-v + w \right)^\top \nabla^2 f(x) \left(-v + w \right) \\ =\ & 4\delta^2 v^\top \nabla^2 f(x) w, \end{aligned} \tag{10}$$

where $\nabla^2 f$ denotes the Hessian of f.

From the above derivation, we see that (9) removes the dependence on the zeroth-order and first-order information, and symmetrizes the estimation (Feng and Wang, 2022). This can reduce variance and stabilize the estimation. A similar phenomenon for the gradient estimators is noted by Duchi et al. (2015).

4.1.1 A Random Projection Derivation

Similar to gradient estimators (Nesterov and Spokoiny, 2017; Li et al., 2020; Wang et al., 2021; Feng and Wang, 2022), one may also derive the Hessian estimator (9) using a random projection argument. Here we use the spherical random projection argument to derive the Hessian estimator. A more thorough study can be found in (Feng and Wang, 2022). To start with, we first prove an identity for random matrix projection in Lemma 3.

Lemma 3. *Let v, w be independently uniformly sampled from the unit sphere in $\mathbb{R}^n$. For any matrix $A \in \mathbb{R}^{n \times n}$, we have*

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

Proof. It is sufficient to show $\mathbb{E}\left[v^i A_i^j w_j v^k w_l \right] = \frac{1}{n^2} A_l^k$ for any $k, l \in [n]$ (Einstein's notation is used).

Since v is uniformly sampled from $\mathbb{S}^{n-1}$ (the unit sphere in $\mathbb{R}^n$), for $k \neq i$, we have $\mathbb{E}\left[v^i v^k | v^k = x \right] = 0$ for any x. This gives that

$$\mathbb{E}\left[v^i v^k \right] \overset{(i)}{=} \int_{x \in [-1,1]} \mathbb{P}\left(v^k = x \right) \mathbb{E}\left[v^i v^k | v^k = x \right] \, dx = 0, \qquad \forall k \neq i.$$

By symmetry of the sphere $\mathbb{S}^{n-1}$ and that $\mathbb{E}\left[v^i v_i\right] = 1$, we have $\mathbb{E}\left[v^k v^k\right] \overset{(ii)}{=} \frac{1}{n}$ for any $k \in [n]$. Combining (i) and (ii) gives

$$\mathbb{E}\left[v^i v^k\right] \overset{(iii)}{=} \frac{1}{n}\delta^{ki},$$

where δ^{ki} is the Kronecker's delta with two superscript.

Similarly, it holds that $\mathbb{E}\left[w_j w_l\right] \overset{(iv)}{=} \frac{1}{n}\delta_{jl}$, where δ_{jl} is the Kronecker's delta with two subscript. Since v and w are independent, (iii) and (iv) gives

$$\mathbb{E}\left[v^i A_i^j w_j v^k w_l\right] = \frac{1}{n^2} A_i^j \delta^{ik}\delta_{jl} = \frac{1}{n^2} A_l^k,$$

which concludes the proof. □

With Lemma 3, we can see that the estimator in (9) satisfies

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\mathrm{H}} f(p; v, w; \delta)\right] \\
&\overset{(i)}{\approx} \frac{n^2}{8\delta^2}\mathbb{E}\left[4\delta^2 \left(v^\top \nabla^2 f(x) w\right)\left(vw^\top + wv^\top\right)\right] \\
&= \frac{n^2}{2}\left(\mathbb{E}\left[\left(v^\top \nabla^2 f(x) w\right) wv^\top\right] + \mathbb{E}\left[\left(w^\top \left[\nabla^2 f(x)\right]^\top v\right) vw^\top\right]\right) \\
&\overset{(ii)}{=} \frac{1}{2}\nabla^2 f(x) + \frac{1}{2}\left[\nabla^2 f(x)\right]^\top \\
&= \nabla^2 f(x),
\end{aligned} \tag{11}$$

where (i) uses (10), and (ii) uses Lemma 3. The above argument gives a random-projection derivation for the estimator (9).

Similar to (Feng and Wang, 2022), we can obtain an $O(\delta^2)$ bias bound in Euclidean spaces.

Corollary 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function, and let $\partial^k f$ ($k \in \mathbb{N}_+$) denote the k-th order total derivative of f. Let f be 4-th order continuously differentiable. Let there be a constant L_4 such that $\|\partial^4 f(x)\| \le L_4$ for all $x \in \mathbb{R}^n$, where $\|\cdot\|$ denotes the spectral norm (∞-Schatten norm) of a tensor. Then it holds that*

$$\left\|\mathbb{E}_{v,w}\widehat{\mathrm{H}} f(p; v, w; \delta) - \mathrm{Hess} f(p)\right\| \le \frac{L_4 n \delta^2}{n+2}, \quad \forall p \in \mathbb{R}^n, \delta \in (0, \infty),$$

where v, w are uniformly sampled from the unit sphere $\mathbb{S}^{n-1}$.

Proof. Firstly, note that the injectivity radius of Euclidean spaces is infinite. By Lemmas 1 and 2, it suffices to consider $\|\mathrm{Hess}\widehat{f}^{\delta}(p) - \mathrm{Hess} f(p)\|$. In the Euclidean case, for any $u, v, p \in \mathbb{R}^n$, Taylor's theorem gives

$$\mathrm{Hess} f(p+v)(u,u) = \mathrm{Hess} f(p)(u,u) + \partial^3 f(p)(u,u,v) + \frac{1}{2}\partial^4 f(p')(u,u,v,v)$$

for some p' depending on p, u, v. Fix any unit vector $u \in \mathbb{R}^n$, we have

$$\begin{aligned}
&\mathrm{Hess}\widetilde{f}^{\delta}(p)(u,u) - \mathrm{Hess} f(p)(u,u) \\
&= \frac{1}{\delta^{2n} V_n^2}\int_{v\in\delta\mathbb{B}^n}\int_{w\in\delta\mathbb{B}^n} \mathrm{Hess} f(p+v+w)(u,u)\, dv\, dw - \mathrm{Hess} f(p)(u,u) \\
&\qquad\qquad (\delta\mathbb{B}^n \text{ is the origin-centered ball with radius } \delta.) \\
&= \frac{1}{\delta^{2n} V_n^2}\int_{v\in\delta\mathbb{B}^n}\int_{w\in\delta\mathbb{B}^n} \left(\partial^3 f(p)(u,u,v+w) + \frac{1}{2}\partial^4 f(p')(u,u,v+w,v+w)\right) dv\, dw \\
&= \frac{1}{2\delta^{2n} V_n^2}\int_{v\in\delta\mathbb{B}^n}\int_{w\in\delta\mathbb{B}^n} \partial^4 f(p')(u,u,v+w,v+w)\, dv\, dw. \qquad \text{(by symmetry of the ball } \delta\mathbb{B}^n)
\end{aligned}$$

By symmetry of $\delta\mathbb{B}^n$, we know that $\int_{v\in\delta\mathbb{B}^n}\int_{w\in\delta\mathbb{B}^n}\partial^4 f(p')(u,u,v,w)\,dv\,dw=0$. Since $\|\partial^4 f(p)\|\le L_4$ for all $p\in\mathbb{R}^n$, we have that

$$\left|\int_{v\in\delta\mathbb{B}^n}\int_{w\in\delta\mathbb{B}^n}\partial^4 f(p')(u,u,v+w,v+w)\,dv\,dw\right|\le L_4\frac{2}{n(n+2)}A_n^2\delta^{2n+2},$$

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

$$\left|\mathrm{Hess}\widetilde{f}^{\delta}(p)(u,u)-\mathrm{Hess}f(p)(u,u)\right|\le\frac{L_4 n\delta^2}{n+2}.$$

We can conclude the proof since the above inequality holds for arbitrary unit vector u. □

4.2 Zeroth Order Hessian Inversion

4.2.1 Hessian Adjugate Estimation by Cramer's Rule

Cramer's rule states that the inverse of a nonsingular matrix A equals

$$A^{-1}=\frac{1}{\det(A)}\mathrm{adj}(A),$$

where $\mathrm{adj}(A)$ is the adjugate of matrix A. Recall the adjugate of matrix A is

$$\mathrm{adj}(A)=\left[(-1)^{i+j}M_{ji}\right]_{\{1\le i,j\le n\}},$$

where $M_{ji}=\det\left(A_{-ji}\right)$ and A_{-ji} is the submatrix of A by removing the j-th row and i-th column. As suggested by the Cramer's rule, one can estimate inverse of Hessian (up to scaling) by first estimating the unsigned minors of the Hessian and then gather the minors into a matrix. This estimation procedure is summarized in Algorithm 1.

Proposition 2. *Let* $\texttt{CHA}(m,\delta,x)$ *be the estimator returned by Algorithm 1. If* $\nabla^2\widetilde{f}^{\delta}(x)$ *is non-singular, it holds that*

$$\mathbb{E}\left[\texttt{CHA}(m,\delta,x)\right]=\det\left(\nabla^2\widetilde{f}^{\delta}(x)\right)\nabla^{-2}\widetilde{f}^{\delta}(x),$$

where $\nabla^{-2}\widetilde{f}^{\delta}(x):=\left[\nabla^2\widetilde{f}^{\delta}(x)\right]^{-1}$.

Proof. We will use the notations defined in Algorithm 1. By Lemma 1, we know that, $\forall i,j,a,b\in[n],\ \ \forall k\in[m]$,

$$\mathbb{E}\left[\widehat{\mathrm{S}}_{k,ij}\right]=\mathbb{E}\left[\widehat{\mathrm{S}}_{k,ij,ab}\right]=\left[\mathbb{E}\left[\widehat{\mathrm{H}}f(x;v_{k,ij,ab},w_{k,ij,ab};\delta)\right]\right]_{-ij}=\left[\nabla^2\widetilde{f}^{\delta}(x)\right]_{-ij}.$$

Since (i) the determinant of a matrix can be expressed in terms of multiplication and addition of its entries, and (ii) all entries of $\widehat{\mathrm{S}}_{k,ij}$ are independent, we have

$$\mathbb{E}\left[\det\left(\widehat{\mathrm{S}}_{k,ij}\right)\right]=\det\left(\mathbb{E}\left[\widehat{\mathrm{S}}_{k,ij}\right]\right).$$

By a use of the Cramer's rule and the above result, it holds that

$$\mathbb{E}\left[\texttt{CHA}(m,\delta,x)\right]=\mathbb{E}\left[(-1)^{i+j}\widehat{M}_{ji}\right]=\mathrm{adj}\left(\nabla^2\widetilde{f}^{\delta}(x)\right)=\det\left(\nabla^2\widetilde{f}^{\delta}(x)\right)\nabla^{-2}\widetilde{f}^{\delta}(x).$$

□

The biggest advantage of the `CHA` method is that it gives an unbiased estimator of the adjugate matrix of $\nabla^2\widetilde{f}^{\delta}(x)$. Also, Proposition 2 hold true even if $\nabla^2\widetilde{f}^{\delta}(x)$ is non-definite. However, a shortcoming of the `CHA` method is its computational expense. For this reason, we introduce the following zeroth order Hessian inversion method, for a special class of Hessian matrices.

Algorithm 1 Cramer-Hessian-Adjugate (`CHA`)

1: **Input:** number of samples m; finite difference step size δ; location for estimation x.

2: Uniformly independently sample $\{(v_{k,ij,ab}, w_{k,ij,ab})\}_{1\le k\le m, 1\le i,j\le n, 1\le a,b\le n}$ from $\mathbb{S}^{n-1}$ (the unit sphere in $\mathbb{S}^{n-1}$).

3: For all $i, j \in [n]$ and $k \in [m]$, create n^2 estimators for the (i,j)-submatrix of $\left[\nabla^2 \widetilde{f}^{\delta}(x)\right]_{-ij}$ by

$$\widehat{\mathrm{S}}_{k,ij,ab} = \left[\widehat{\mathrm{H}} f(x; v_{k,ij,ab}, w_{k,ij,ab}; \delta)\right]_{-ij}, \quad \forall 1 \le i,j,a,b \le n, \ \ \forall k \in [m].$$

4: Create estimators of $\left[\nabla^2 \widetilde{f}^{\delta}(x)\right]_{-ij}$ (written $\widehat{\mathrm{S}}_{k,ij}$) such that the (a,b)-th entry of $\widehat{\mathrm{S}}_{k,ij}$ is the (a,b)-th entry of $\widehat{\mathrm{S}}_{k,ij,ab}$ for all $a, b \in [n]$.
/* In practice, one can use the entry-wise estimators to replace the estimator in Step 3. See Section 5.1.1 for more details on entry-wise Hessian estimators. */

5: For all $i, j \in [n]$, estimate the unsigned minors by

$$\widehat{M}_{ij} = \frac{1}{m} \sum_{k=1}^{m} \det\left(\widehat{\mathrm{S}}_{k,ij}\right).$$

/* The determinant can be computed via LU decomposition, QR decomposition, or similar methods. */

6: Estimate the adjugate of Hessian by

$$\overline{\mathrm{A}}_m \widetilde{f}^{\delta}(x) = \left[(-1)^{i+j} \widehat{M}_{ji}\right]. \tag{12}$$

7: **Output:** $\texttt{CHA}(m, \delta, x) = \overline{\mathrm{A}}_m \widetilde{f}^{\delta}(x)$.

4.2.2 Hessian Inverse Estimation by Neumman Series

An approach for computing the inverse of Hessian is via Neumman series. For an invertible matrix A satisfying $\lim_{p\to\infty} (I - A)^p = 0$, the Neumann series expands the inverse of A by

$$A^{-1} = \sum_{p=0}^{\infty} (I - A)^{-1}.$$

From this observation, we can first estimate the Hessian, and then estimate the inverse by the Neumann series. Previously, Agarwal et al. (2017) studied fast Neumann series based Hessian inversion using first-order information. Here a similar result can be obtained using zeroth-order information only. This zeroth-order extension of (Agarwal et al., 2017) is summarized in Algorithm 2.

Algorithm 2 Neumman-Hessian-Inverse (`NHI`)

1: **Input:** number of samples (m_1, m_2, m_3); finite difference step size δ; location for estimation x.
2: Uniformly independently sample $\{(v_{ijk}, w_{ijk})\}_{1\le i\le m_1, 1\le j\le m_2, 1\le k\le m_3}$ from $\mathbb{S}^{n-1}$.
3: For all i, j, k, compute

$$\widehat{\mathrm{H}} f(x; v_{ijk}, w_{ijk}; \delta),$$

as defined in (9).
4: Compute

$$\overline{\mathrm{H}}^{-1}_{m_1,m_2,m_3} \widetilde{f}^{\delta}(x) = \frac{1}{m_1}\sum_{i=1}^{m_1}\left(I + \sum_{h=1}^{m_2}\prod_{j=1}^{h}\left(I - \frac{1}{m_3}\sum_{k=1}^{m_3}\widehat{\mathrm{H}} f(x; v_{ijk}, w_{ijk}; \delta)\right)\right). \tag{13}$$

5: **Output:** $\texttt{NHI}(m_1, m_2, m_3, \delta, x) = \overline{\mathrm{H}}^{-1}_{m_1,m_2,m_3} \widetilde{f}^{\delta}(x)$.

Proposition 3. *Suppose f is twice-differentiable, α-strongly convex and β-smooth with $\beta < 1$. Then it holds that*

$$\left\| \mathbb{E}\left[\textit{NHI}(m_1, m_2, m_3, \delta, x)\right] - \nabla^{-2}\widetilde{f}^{\delta}(x) \right\| \le \frac{(1-\alpha)^{m_2+1}}{\alpha}, \tag{14}$$

where $\nabla^{-2}\widetilde{f}^{\delta}(x) := \left[\nabla^2 \widetilde{f}^{\delta}(x)\right]^{-1}$.

Proof. Since f is α-strongly convex, it holds that, for any $x, y, v, w \in \mathbb{R}^n$,

$$f(x+v+w) \ge f(y+v+w) + (x-y)^{\top}\nabla f(y+v+w) + \frac{\alpha}{2}\|x-y\|^2.$$

Integrating both v and w over $\delta\mathbb{B}^n$ gives that

$$\widetilde{f}^{\delta}(x) \ge \widetilde{f}^{\delta}(y) + (x-y)^{\top}\nabla\widetilde{f}^{\delta}(y) + \frac{\alpha}{2}\|x-y\|^2,$$

where we use the dominated convergence theorem to interchange the integral and the gradient operator. This shows that $\widetilde{f}^{\delta}$ is also α-strongly.

Since a differentiable function f is β-smooth if and only if $f(x) \le f(y) + \nabla f(y)^{\top}(x-y) + \frac{\beta}{2}\|x-y\|^2$ for all $x, y \in \mathbb{R}^n$, one can show that $\widetilde{f}^{\delta}$ is β-smooth by repeating the above argument.

For $\texttt{NHI}(m_1, m_2, m_3, \delta, x)$, we have

$$\begin{aligned}\mathbb{E}\left[\texttt{NHI}(m_1, m_2, m_3, \delta, x)\right] &= I + \sum_{h=1}^{m_2}\prod_{j=1}^{h}\left(I - \mathbb{E}\left[\widehat{\mathrm{H}} f(x; v_{ijk}, w_{ijk}; \delta)\right]\right)\\ &= \sum_{j=0}^{m_2}\left(I - \nabla^2\widetilde{f}^{\delta}(x)\right)^j.\end{aligned}$$

Since $\widetilde{f}^{\delta}$ is α-strongly convex, β-smooth ($\beta < 1$), and apparently twice-differentiable, we have

$$0 \preccurlyeq I - \nabla^2\widetilde{f}^{\delta}(x) \preccurlyeq (1-\alpha)\, I.$$

Thus we can bound the bias by

$$\left\| \mathbb{E}\left[\texttt{NHI}(m_1, m_2, m_3, \delta, x)\right] - \nabla^{-2}\widetilde{f}^{\delta}(x) \right\| \le \sum_{j=m_2+1}^{\infty}(1-\alpha)^j = \frac{(1-\alpha)^{m_2+1}}{\alpha}.$$

□

5 Existing Methods and Experiments

5.1 Existing Methods for Hessian Estimation

5.1.1 Hessian Estimation via Collecting Single Entry Estimations

In the Euclidean case, one can fix a canonical coordinate system $\{\boldsymbol{e}_i\}_{i\in[n]}$, and the (i,j)-th entry of the Hessian matrix of f at x can be estimated by

$$\begin{aligned}\widehat{\mathrm{H}}_{ij}^{\mathrm{entry}} f(x;\delta)= \frac{1}{4\delta^2}\big(&f(x+\delta\boldsymbol{e}_i+\delta\boldsymbol{e}_j)-f(x+\delta\boldsymbol{e}_i-\delta\boldsymbol{e}_j)\\ &-f(x-\delta\boldsymbol{e}_i+\delta\boldsymbol{e}_j)+f(x-\delta\boldsymbol{e}_i-\delta\boldsymbol{e}_j)\big).\end{aligned} \tag{15}$$

One can then gather the entries to obtain a Hessian estimator:

$$\widehat{\mathrm{H}}^{\mathrm{entry}} f(x;\delta) = \left[\widehat{\mathrm{H}}_{ij}^{\mathrm{entry}} f(x;\delta)\right]_{i,j\in[n]}. \tag{16}$$

This estimator could perhaps date back to classic times when the finite difference principles were first used. Yet it needs at least $\Omega(n^2)$ zeroth order samples to produce an estimator in an n-dimensional space. Previously, Balasubramanian and Ghadimi (2021) designed a Hessian estimator based on the Stein's identity (Stein, 1981). Their estimator only needs $O(1)$ zeroth-order function evaluations. This method is discussed in the next section.

5.1.2 Hessian Estimation via the Stein's identity

A classic result for Hessian computation is the Stein's identity (named after Charles Stein), as stated below.

Theorem 2 (Stein's identity)**.** *Consider a smooth function $f:\mathbb{R}^n\to\mathbb{R}$. For any point $x\in\mathbb{R}^n$, it holds that*

$$\nabla^2 f(x) = \frac{1}{2}\mathbb{E}\left[\left(uu^\top - I\right) D_{uu} f(x)\right],$$

where 1. $u\sim\mathcal{N}(0,I)$, and 2.

$$D_{uu} f(x) = \lim_{\tau\to 0}\frac{f(x+\tau u)-2f(x)+f(x-\tau u)}{\tau^2}.$$

Proof. For completeness, a convenient proof of Theorem 2 is provided in the Appendix. □

One can estimate Hessian using the Stein's identity (Balasubramanian and Ghadimi, 2021):

$$\widehat{\mathrm{H}}^{\mathrm{Stein}} f(x;u;\delta) = \frac{f(x+\delta u)-2f(x)+f(x-\delta u)}{2\delta^2}(uu^\top - I), \tag{17}$$

where $u\sim\mathcal{N}(0,I)$ is a standard Gaussian vector. A bias bound for (17) is in Theorem 3.

Theorem 3 (Li et al. (2020); Balasubramanian and Ghadimi (2021))**.** *Let f have L_2-Lipschitz Hessian: There exists a constant L_2 such that $\|\nabla^2 f(x)-\nabla^2 f(x')\|\le L_2\|x-x'\|$ for all $x,x'\in\mathbb{R}^n$. The estimator in (17) satisfies*

$$\left\|\mathbb{E}\left[\widehat{\mathrm{H}}^{\mathrm{Stein}} f(x;u;\delta)\right]-\nabla^2 f(x)\right\| \le \frac{L_2\,(n+6)^{\frac{5}{2}}\,\delta}{4},$$

for any $x\in\mathbb{R}^n$ and any function $f:\mathbb{R}^n\to\mathbb{R}$ with L_2-Lipschitz Hessian.

The estimator (17) improves the entry-wise estimator in the sense that one only needs $O(1)$ samples to produce an estimator. However, its error bound given by Theorem 3 is worse than that of (9) in Theorem 1. A more detailed discussion on the error bounds is in Remark 1.

Remark 1. *We need to note that our estimator (9) and the estimator via Stein's method (17) have different finite-difference step size. More specifically,* $\mathbb{E}_{v,w \overset{i.i.d.}{\sim} \mathbb{S}^{n-1}} [\delta \|v+w\|] = \Theta(\delta)$ *for (9) and* $\mathbb{E}_{u\sim\mathcal{N}(0,I_n)} [\delta\|u\|] = \Theta(\sqrt{n}\delta)$ *for (17). To compare the bias bounds for (9) and (17) using the same (expected) finite-difference step size, we need to downscale the bound in Theorem 3 by a factor of* $\sqrt{n}$*. After this downscaling, the error bound for the Stein-type estimator (17) is* $O\left(L_2 n^2 \delta\right)$ *which is worse than the bias bound bound for our estimator (9). In the experiments, we down scale the finite-difference step size when studying all results of Stein's method estimator.*

As discussed in Remark 1, a difference between (9) and (17) is that they sample random vectors from different distributions: uniformly random vector from the unit sphere for (9) and standard Gaussian vector for (17). High moments of uniformly random vectors from the unit sphere are smaller than Gaussian vectors of same expected norm. More specifically, The k-th moment of (norm of) a standard Gaussian vector $v \sim \mathcal{N}(0, I_n)$ that is downscale by a factor of $\sqrt{n}$ is

$$
\begin{aligned}
& n^{-k/2}\mathbb{E}\left[\|v\|^k\right] \\
&= \frac{n^{-k/2}}{(2\pi)^{-n/2}} \int_0^{\pi}\int_0^{\pi}\cdots\int_0^{2\pi}\int_0^{\infty} r^{k+n-1}e^{-\frac{r^2}{2}}\, dr\, \sin^{n-2}(\varphi_1)\sin^{n-3}(\varphi_2)\cdots\sin(\varphi_{n-2})\, d\varphi_1\, d\varphi_2\cdots d\varphi_{n-1} \\
&= n^{-k/2}\,(2\pi)^{-n/2}\, A_n \int_0^{\infty} r^{k+n-1}e^{-\frac{r^2}{2}}\, dr \\
& \qquad\qquad (A_n \text{ is the surface area of the Euclidean unit sphere } \mathbb{S}^{n-1}) \\
&= n^{-k/2}\,(2\pi)^{-n/2}\,\frac{2\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} 2^{\frac{n+k}{2}-1}\Gamma\left(\frac{k+n}{2}\right) \\
&= n^{-k/2} 2^{\frac{k}{2}} \frac{\Gamma\left(\frac{k+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \\
&\sim n^{-k/2} 2^{\frac{k}{2}} \frac{\sqrt{\frac{2\pi}{\frac{k+n}{2}}}\left(\frac{\frac{k+n}{2}}{e}\right)^{\frac{k+n}{2}}}{\sqrt{\frac{2\pi}{\frac{n}{2}}}\left(\frac{\frac{n}{2}}{e}\right)^{\frac{n}{2}}} \qquad \text{(by Stirling's approximation)} \\
&= \left(\frac{en}{2}\right)^{-k/2} \sqrt{\frac{n}{n+k}} \left(\frac{n+k}{2}\right)^{n/2+k/2} \left(\frac{n}{2}\right)^{-n/2}
\end{aligned}
$$

which clearly grows very fast with k for large k and for any fixed n. On contrary, the moments of (norm) of the vector uniformly sampled from the unit sphere are all 1. This difference implies that our estimator tends to have smaller higher order moments.

5.2 Numerical Results

We test the performance of our estimator against the previous two estimators in noisy environments. Before proceeding, we re-define some notations for the estimators, so that the estimators are tested on the same ground and noise is properly taken into consideration. The estimators we will empirically study are

1. Our new estimator:

$$
\begin{aligned}
& \widehat{\mathrm{H}}^{\mathrm{new}} f(p; m; \delta) \\
&= \sum_{k=1}^{\lfloor \frac{m}{4} \rfloor} \frac{n^2}{\delta^2} \big[\epsilon_k + f(\mathrm{Exp}_p(\delta v_k + \delta w_k)) - f(\mathrm{Exp}_p(-\delta v_k + \delta w_k)) \\
&\quad - f(\mathrm{Exp}_p(\delta v_k - \delta w_k)) + f(\mathrm{Exp}_p(-\delta v_k - \delta w_k))\big](v_k \otimes w_k + w_k \otimes v_k), \qquad (18)
\end{aligned}
$$

where $v_k, w_k \overset{i.i.d.}{\sim} \mathbb{S}_p$, and ϵ_k is a mean-zero noise that is independent of all other randomness.

2. The Stein's estimator:

$$\begin{aligned}&\widehat{\mathrm{H}}^{\mathrm{Stein}} f(p; m; \delta)\\ =& \sum_{k=1}^{\lfloor \frac{m}{3} \rfloor} \frac{n}{2\delta^2} \left[f\left(\mathrm{Exp}_p\left(\frac{\delta u_k}{\sqrt{n}}\right)\right) - 2f(p) + f\left(\mathrm{Exp}_p\left(\frac{-\delta u_k}{\sqrt{n}}\right)\right) + \epsilon_k \right]\\ &\cdot (u_k \otimes u_k - I),\end{aligned} \tag{19}$$

where $u_k \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ (the standard Gaussian in $T_p\mathcal{M}$), I is the identity map from $T_p\mathcal{M}$ to itself (As a bilinear form, $I(u, v) = \langle u, v \rangle_p$ for any $u, v \in T_p\mathcal{M}$.), and ϵ_k is a mean-zero noise that is independent of all other randomness.

3. The entry-wise estimator:

$$\widehat{\mathrm{H}}^{\mathrm{entry}} f(p; m; \delta) = \left[\widehat{\mathrm{H}}_{ij}^{\mathrm{entry}} f(p; m; \delta) \right]_{i,j \in [n]}, \tag{20}$$

where

$$\begin{aligned}\widehat{\mathrm{H}}_{ij}^{\mathrm{entry}} f(p; m; \delta) = \frac{1}{4\delta^2} \sum_{k=1}^{\lfloor \frac{m}{4n^2} \rfloor} &\big(f(\mathrm{Exp}_p(\delta \boldsymbol{e}_i + \delta \boldsymbol{e}_j)) - f(\mathrm{Exp}_p(\delta \boldsymbol{e}_i - \delta \boldsymbol{e}_j))\\ &- f(\mathrm{Exp}_p(-\delta \boldsymbol{e}_i + \delta \boldsymbol{e}_j)) + f(\mathrm{Exp}_p(-\delta \boldsymbol{e}_i - \delta \boldsymbol{e}_j)) + \epsilon_k \big),\end{aligned}$$

$\{\boldsymbol{e}_i\}_i$ is the local normal coordinate for $T_p\mathcal{M}$, and ϵ_k is a mean-zero noise that is independent of all other randomness.

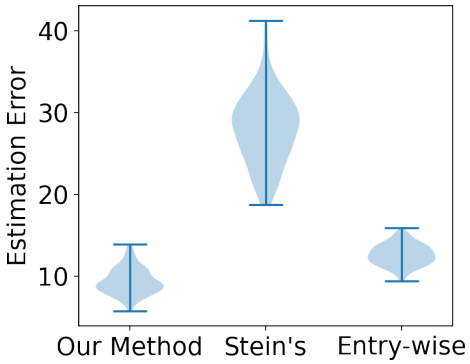
Strictly speaking, the noises ϵ_k corrupt all the zeroth-order function value observations. Specifically, the notation $\epsilon_k + f(\mathrm{Exp}_p(\delta v_k + \delta w_k)) - f(\mathrm{Exp}_p(-\delta v_k + \delta w_k)) - f(\mathrm{Exp}_p(\delta v_k - \delta w_k)) + f(\mathrm{Exp}_p(-\delta v_k - \delta w_k))$ should be understood in the following way. All four functions values $f(\mathrm{Exp}_p(\delta v_k + \delta w_k))$, $f(\mathrm{Exp}_p(-\delta v_k + \delta w_k))$, $f(\mathrm{Exp}_p(\delta v_k - \delta w_k))$, and $f(\mathrm{Exp}_p(-\delta v_k - \delta w_k))$ are corrupted with mean-zero and independent noise and not directly observable. Note that all previously discussed bias bounds hold when the function evaluations are corrupted by independent mean-zero noise.

Table 1: Manifolds used for testing. The local metric near p is implicitly specified by the exponential map.

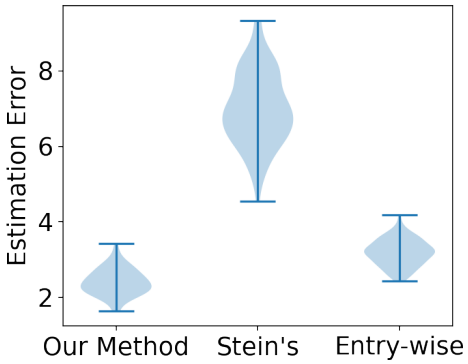
Manifold	p ($p \in \mathbb{R}^{n+1}$)	$h(x)$, $x \in T_p\mathcal{M} \cong \mathbb{R}^n$	$\mathrm{Exp}_p(v)$
(I)	$p = 0$	$h(x) = 0$	$(v, h(v))$
(II)	$p = 0$	$h(x) = 1 - \sqrt{1 - \sum_{i=1}^n x_i^2}$	$(v, h(v))$
(III)	$p = 0$	$h(x) = \sum_{i=1}^{n/2} x_i^2 - \sum_{i=n/2+1}^n x_i^2$	$(v, h(v))$

Table 2: Timing results in seconds, rounded to 1e-4 accuracy. In the table, "Our method" stands for the estimator (18), and "Stein's" stands for the estimator (19). The time consumption of the estimators are divided into three parts: (1) sampling time, used for generating the random vectors (uniformly random unit vectors for our methods, and standard Gaussian vectors for the Stein's method), (2) evaluation time, used for accessing function values, and (3) computation time, used for matrix manipulations (e.g., outer product computation). In the timing experiments, both estimators (18) and (19) use $m = 10,000$, $\delta = 0.05$ and $n = 8$. All timing results are averaged 10 times, presented in a "mean ± standard deviation" format. The last two columns are cumulative, to avoid fast memory access to saved data.

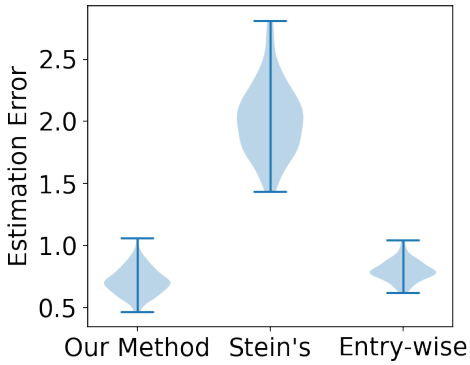
	Sampling	Sampling + Evaluation	Sampling + Evaluation + Computation
Our method	0.1580 ± 0.0031	0.5763 ± 0.0036	0.6977 ± 0.0040
Stein's	0.0257 ± 0.0012	0.3020 ± 0.0024	0.4222 ± 0.0028



(a) $\delta = 0.05$



(b) $\delta = 0.1$



(c) $\delta = 0.2$

Figure 1: Results for the Manifold (I). Each violin plot summarizes estimation error of 100 estimations. More specifically, each estimation in this figure uses $m = 3840$ function evaluations, and 100 estimations are used to generate one violin plot. On the x-axis, "Our method" corresponds to our estimator (18); "Stein's" corresponds to the Stein's method (19); "Entry-wise" corresponds to the entry-wise estimator (20). Subfigures (a), (b), (c) corresponds to $\delta = 0.05$, $\delta = 0.1$, $\delta = 0.2$.

The above notations allow us to put all the estimators on the same ground more easily. With the new notations, all $\widehat{\mathrm{H}}^{\mathrm{new}} f(p; m; \delta)$, $\widehat{\mathrm{H}}^{\mathrm{Stein}} f(p; m; \delta)$ and $\widehat{\mathrm{H}}^{\mathrm{entry}} f(p; m; \delta)$ uses m functions value observations and have an expected finite difference step size $\Theta(\delta)$. The redefining of the estimators is needed since 1. the entry-wise estimator needs more samples to output an estimate, and 2. the default Stein's method in expectation uses a larger finite-difference step-size, as discussed in Remark 1.

Remark 2. *The estimator we introduced (18) has a practical advantage over that via the Stein's identity (19). The reason is that estimators based on the Stein's identity requires one to explicitly know the identity map from $T_p\mathcal{M}$ to itself. This map may or may not admit an easy numerical representation. For example, for the real Stiefel's manifold $\mathrm{St}(n, k) = \{X \in \mathbb{R}^{n\times k} : X^\top X = I\}$, we know that the map*

$$P_X Z := (I - XX^\top)Z + \frac{1}{2}X\left(X^\top Z - Z^\top X\right), \qquad \forall Z \in \mathbb{R}^{n\times k},$$

is the identity from $T_X\mathrm{St}(n, k)$ to itself (e.g., Absil et al., 2009). Also, this map projects any $Z \in \mathbb{R}^{n\times k}$ onto $T_X\mathrm{St}(n, k)$. Extracting a numerical representation of this map P_X may not be easy. On contrary, for any $Z_1, Z_2 \in T_X\mathrm{St}(n, k)$, computing $Z_1 \otimes Z_2$ is tractable. More specifically, one can use the following procedure to obtain a uniformly random vector from the unit sphere in

$T_X\mathrm{St}(n,k)$ for a given $X \in \mathrm{St}(n,k)$. One can (1) sample an i.i.d. Gaussian matrix G from $\mathbb{R}^{n,k}$, (2) compute $P_X G$, and (3) normalize $P_X G$ with respect to the Frobenius inner product. By rotational invariance (of the standard Gaussian distribution and the Frobenius norm), this procedure outputs a uniformly random unit matrix in $T_X\mathrm{St}(n,k)$. Once we have the unit vectors in $T_X\mathrm{St}(n,k)$, we can numerically compute their tensor product.

All three methods are tested using the following test function, defined using the standard Cartesian coordinate system in $\mathbb{R}^{n+1}$,

$$f(x) = \sum_{i=1}^{n+1} \cos(x_i) + \exp(x_1 x_2).$$

Every function evaluation is corrupted with an independent noise sampled from $\mathcal{N}(0, 0.0025)$. The estimators are tested over three manifolds in $\mathbb{R}^{n+1}$. More details about the three manifolds are in Table 1. In all settings, we set the number of total function evaluation $m = 3840$ and dimension of manifold $n = 8$. The results for manifold (I), the Euclidean case, is in Figure 1. Results for manifold (II) and manifold (III) are in Appendix B. In Figure 1 (and Figures 2 and 3 in Appendix B), the errors on the y-axis plots the norm of the difference between the empirical estimator and the truth:

$$\left\| \widehat{\mathrm{H}} f(p; v, w; \delta) - \mathrm{Hess} f(p) \right\|.$$

5.3 Time Efficiency Comparison

We compare the time efficiency of our method and the estimator based on the Stein's identity. In general, one expects estimators based on the Stein's identity to be more time-efficient. Main reasons for this include that the estimator based on the Stein's identity needs only 3 function evaluations instead of 4. In practice, the function evaluations may or may not be cheap. When the functions evaluations are expensive, we may expect that estimators based on the Stein's identity approximately saves $1/4$ time, compared to our method. When the function evaluations are cheap, our estimator (18) is in general more time consuming as well, since more sampling and more matrix computations need to be carried out.

In Table 2, we compare the running time of (18) and (19). All timing experiments use the same benchmark function and underlying manifold as Figure 1. We use $n = 8$, $m = 10,000$ and $\delta = 0.05$ for timing experiments. All timing experiments are carried out in an environment with

- 10 cores and 20 logical processors with a maximum speed of 2.80 GHz;
- 32GB RAM;
- Windows 11 22000.832;
- Python 3.8.8.

6 Conclusion

In this paper, we study Hessian estimators over Riemannian manifolds. We design a new estimator, such that for real-valued analytic functions over an n-dimensional complete analytic Riemannian manifold, our estimator achieves an $O(\gamma\delta^2)$ expected error, where γ depends both on the Levi-Civita connection and the function f, and δ is the finite difference step size. Downstream computations of Hessian inversion is also studied. Empirical studies show supremacy of our method over existing methods.

Data Availability Statement

No new data were generated or analysed in support of this paper. Code for the experiments is available at `https://github.com/wangt1anyu/zeroth-order-Riemann-Hess-code`.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Agarwal, N., Bullins, B., and Hazan, E. (2017). Second-order stochastic optimization for machine learning in linear time. *The Journal of Machine Learning Research*, 18(1):4148–4187.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer.

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.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer.

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.

Feng, Y. and Wang, T. (2022). Stochastic Zeroth Order Gradient and Hessian Estimators: Variance Reduction and Refined Bias Bounds. *arXiv preprint arXiv:2205.14737*.

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.

Gavrilov, A. V. (2007). The double exponential map and covariant derivation. *Siberian Mathematical Journal*, 48(1):56–61.

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

Greene, R. E. and Wu, H. (1979). c^∞-approximations of convex, subharmonic, and plurisubharmonic functions. In *Annales Scientifiques de l'École Normale Supérieure*, volume 12, pages 47–84.

Greene, R. E. and Wu, H.-H. (1973). On the subharmonicity and plurisubharmonicity of geodesically convex functions. *Indiana University Mathematics Journal*, 22(7):641–653.

Greene, R. E. and Wu, H.-h. (1976). c^∞ convex functions and manifolds of positive curvature. *Acta Mathematica*, 137(1):209–245.

Li, J., Balasubramanian, K., and Ma, S. (2020). Stochastic zeroth-order riemannian derivative estimation and optimization. *arXiv preprint arXiv:2003.11238*.

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

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.

Petersen, P. (2006). *Riemannian geometry*, volume 171. Springer.

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.

Spall, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE transactions on automatic control*, 45(10):1839–1853.

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

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 Proof of Theorem 2

Proof of Theorem 2. Consider $\mathbb{E}\left[u_k u_h u_i u_j \partial_i \partial_j\right]$ for any $k, h, i, j \in [n]$.

When $(k,h) = (i,j)$, one has $\mathbb{E}\left[u_k u_h u_i u_j \partial_i \partial_j\right] = \mathbb{E}\left[u_i^2 u_j^2 \partial_i \partial_j\right]$. In this case, it holds that

$$\mathbb{E}\left[u_i^2 u_j^2 \partial_i \partial_j\right] = \partial_k \partial_h \quad \text{for } i \neq j \quad \text{and} \quad \mathbb{E}\left[u_i^4 \partial_i \partial_i\right] = 3\partial_k \partial_k, \quad \text{for } i = j.$$

When $(k,h) \neq (i,j)$, $i = j$ and $k = h$, we have $\mathbb{E}\left[u_k^2 u_i^2 \partial_i \partial_j\right] = \partial_i \partial_i$.

When $(k,h) \neq (i,j)$, $i = j$ and $k \neq h$, we have $\mathbb{E}\left[u_k u_h u_i u_j\right] = 0$.

When $(k,h) \neq (i,j)$, $i \neq j$ and $k = h$, we have $\mathbb{E}\left[u_k u_h u_i u_j\right] = 0$.

When $(k,h) \neq (i,j)$, $i \neq j$, $k \neq h$, $k = j$ and $h = i$, we have $\mathbb{E}\left[u_k u_h u_i u_j \partial_i \partial_j\right] = \mathbb{E}\left[u_i^2 u_j^2 \partial_h \partial_k\right] = \partial_h \partial_k$.

When $(k,h) \neq (i,j)$, $i \neq j$, $k \neq h$, $k = i$ and $h = j$, we have $\mathbb{E}\left[u_k u_h u_i u_j \partial_i \partial_j\right] = \mathbb{E}\left[u_i^2 u_j^2 \partial_k \partial_h\right] = \partial_k \partial_h$.

When $(k,h) \neq (i,j)$, $i \neq j$, $k \neq h$ and $k \neq j$, we have $\mathbb{E}\left[u_k u_h u_i u_j\right] = 0$.

When $(k,h) \neq (i,j)$, $i \neq j$, $k \neq h$ and $h \neq i$, we have $\mathbb{E}\left[u_k u_h u_i u_j\right] = 0$.

Now using Einstein's notation and combining all above cases give

$$\begin{aligned}\mathbb{E}\left[u^k u_h u^i u_j \partial_i \partial^j\right] &= \partial^k \partial_h (1 - \delta_k^h) + \partial^h \partial_k (1 - \delta_h^k) + \delta_k^h \partial_i \partial^i + 2\partial^k \partial_h \delta_k^h \\ &\overset{(i)}{=} 2\partial^k \partial_h + \delta_k^h \partial_i \partial^i,\end{aligned}$$

where δ_k^h is the Kronecker's delta.

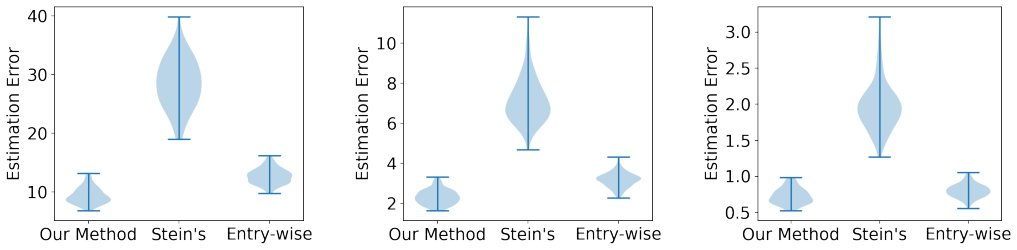
Since $D_{uu} f(p) = u^i u_j \partial_i \partial^j f(x)$ for all u and x, we can write $uu^\top D_{uu} f(x) = u^k u_h u^i u_j \partial_i \partial^j f(x)$. Thus rearranging terms in (i) gives

$$\mathbb{E}\left[uu^\top D_{uu} f(x)\right] \overset{(ii)}{=} 2\nabla^2 f(x) + (\Delta f(x))\, I,$$

where $\Delta = \partial_i \partial^i$ is the Laplace operator.

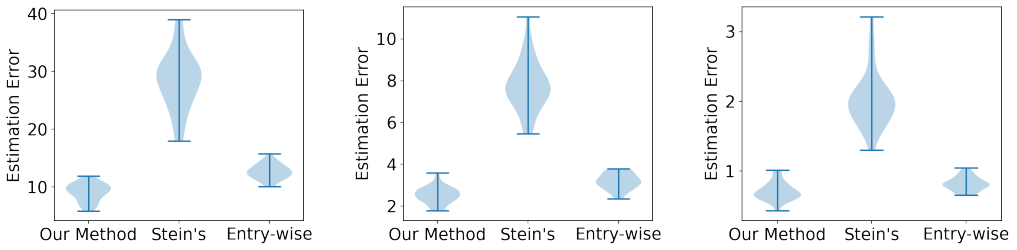
Since $\mathbb{E}\left[D_{uu} f(x)\right] = \mathbb{E}\left[u^i u_j \partial_i \partial^j f(x)\right] = \delta_i^j \partial_i \partial^j f(x) = (\Delta f(x)) I$, rearranging terms in (ii) concludes the proof. □

B Additional Experimental Results



(a) $\delta = 0.05$ (b) $\delta = 0.1$ (c) $\delta = 0.2$

Figure 2: Results for the Manifold (II). Each violin plot summarizes estimation error of 100 estimations. Specifically, each estimation in this figure uses $m = 3840$ function evaluations, and 100 estimations are used to generate one violin plot. On the x-axis, "Our method" corresponds to our estimator (18); "Stein's" corresponds to the Stein's method (19); "Entry-wise" corresponds to the entry-wise estimator (20). Subfigures (a), (b), (c) corresponds to $\delta = 0.05$, $\delta = 0.1$, $\delta = 0.2$.



(a) $\delta = 0.05$ (b) $\delta = 0.1$ (c) $\delta = 0.2$

Figure 3: Results for the Manifold (III). Each violin plot summarizes estimation error of 100 estimations, with the estimators defined in. Specifically, each estimation in this figure uses $m = 3840$ function evaluations, and 100 estimations are used to generate one violin plot. On the x-axis, "Our method" corresponds to our estimator (18); "Stein's" corresponds to the Stein's method (19); "Entry-wise" corresponds to the entry-wise estimator (20). Subfigures (a), (b), (c) corresponds to $\delta = 0.05$, $\delta = 0.1$, $\delta = 0.2$.