On the equivalence of a Hessian-free inequality and Lipschitz continuous Hessian

Radu I. Boţ, Minh N. Dao, Tianxiang Liu, Bruno F. Lourenço, and Naoki Marumo

Abstract It is known that if a twice differentiable function has a Lipschitz continuous Hessian, then its gradients satisfy a Jensen-type inequality. In particular, this inequality is Hessian-free in the sense that the Hessian does not actually appear in the inequality. In this paper, we show that the converse holds in a generalized setting: if a continuous function from a Hilbert space to a reflexive Banach space satisfies such an inequality, then it is Fréchet differentiable and its derivative is Lipschitz continuous. Our proof relies on the Baillon–Haddad theorem.

1 Introduction

Carmon et al. [3] proposed a first-order method for minimizing nonconvex functions having Lipschitz continuous gradients and Hessians. The idea is that even if the Hessian is not actually used, its mere Lipschitz continuity makes it possible to design

Radu I. Boţ
University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria, e-mail: radu.bot@univie.ac.at

Minh N. Dao
School of Science, RMIT University, Melbourne, VIC 3000, Australia, e-mail: minh.dao@rmit.edu.au

Tianxiang Liu
Institute of Systems and Information Engineering, Tsukuba University, Japan, e-mail: liutx@sk.tsukuba.ac.jp

Bruno F. Lourenço
Department of Statistical Inference and Mathematics, Institute of Statistical Mathematics, Japan, e-mail: bruno@ism.ac.jp

Naoki Marumo
Graduate School of Information Science and Technology, University of Tokyo, Japan, e-mail: marumo@mist.i.u-tokyo.ac.jp

faster first-order methods than what would be possible if the function only had Lipschitz continuous gradients. In this vein, several algorithms have been proposed that offer theoretical or practical improvements (e.g., [1, 5, 6, 7, 8, 10]). Among these, Marumo and Takeda [7] proposed a first-order method that does not require the Lipschitz constant of the Hessian as an input of the algorithm, unlike previous methods. An important step in their analysis is establishing the following Jensen-type inequality.

Lemma 1 (Hessian-free inequality [7, Lemma 3.1]). *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function with L-Lipschitz continuous Hessian. Then, for any $x_1,\dots,x_n \in \mathbb{R}^d$ and $\lambda_1,\dots,\lambda_n \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$, we have*

$$\left\| \nabla f\left(\sum_{i=1}^n \lambda_i x_i\right) - \sum_{i=1}^n \lambda_i \nabla f(x_i) \right\| \leq \frac{L}{2} \sum_{1\leq i<j\leq n} \lambda_i \lambda_j \|x_i - x_j\|^2. \tag{1}$$

The proof of Lemma 1, naturally, uses the fact that f has a Lipschitz continuous Hessian. However, the resulting inequality (1) is free of Hessians and only involves the gradient of f.

A natural question then is the following:

> If f is continuously differentiable and satisfies (1) for some constant $L > 0$, is f necessarily twice differentiable? If so, is its Hessian L-Lipschitz continuous?

It turns out that as a consequence of the Baillon–Haddad Theorem the answer is yes. This is somewhat surprising given the fact that (1) does not involve Hessians at all. In fact, we will prove the following Theorem 1, a generalization of the answer to the Hilbert space setting.

In what follows, given Banach spaces X and Y we denote the space of bounded linear operators between X and Y by $\mathscr{B}(X,Y)$. Let X^* denote the dual space of X, i.e., $X^* := \mathscr{B}(X,\mathbb{R})$. Also, for simplicity, throughout the paper we use the same notation $\|\cdot\|$ to indicate the norms on different Banach spaces. We recall that operator norm is defined for $T \in \mathscr{B}(X,Y)$ by $\|T\| := \sup_{x\in X,\,\|x\|\leq 1} \|T(x)\|$.

Theorem 1. *Let X and Y be real Hilbert and reflexive Banach spaces, respectively. Let $F\colon X \to Y$ be a continuous function and let $L > 0$. Then the following are equivalent:*

(i) F is Fréchet differentiable on X and its derivative $F'\colon X \to \mathscr{B}(X,Y)$ is L-Lipschitz continuous, i.e.,

$$\|F'(x) - F'(y)\| \leq L\|x - y\| \quad \forall x, y \in X,$$

where the norm on the left-hand side is the operator norm.

(ii) For any $x_1,\dots,x_n \in X$ and $\lambda_1,\dots,\lambda_n \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$, the following holds:

$$\left\| F\left(\sum_{i=1}^n \lambda_i x_i\right) - \sum_{i=1}^n \lambda_i F(x_i) \right\| \leq \frac{L}{2} \sum_{1\leq i<j\leq n} \lambda_i \lambda_j \|x_i - x_j\|^2. \tag{2}$$

The proof of (i) $\implies$ (ii) is essentially the same as that of Lemma 1, since the proof in [7] does not rely on the assumptions $X = Y = \mathbb{R}^d$ and $F = \nabla f$. In this paper, we focus on the converse implication (ii) $\implies$ (i).

Setting $X = Y = \mathbb{R}^d$ and $F = \nabla f$ establishes the converse of Lemma 1. Slightly more generally, the following result holds.

Corollary 1. *Let X be a real Hilbert space and let $f\colon X \to \mathbb{R}$ be a Fréchet differentiable function such that its gradient ∇f satisfies (1) for every $x_1, \dots, x_n \in X$ and $\lambda_1, \dots, \lambda_n \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$. Then, f is twice differentiable with L-Lipschitz continuous Hessian.*

2 Proof of the Theorem

We start with a result that is contained in the enhanced version of the Baillon–Haddad Theorem described by Bauschke and Combettes in [2].

Theorem 2 (A piece of the Baillon–Haddad Theorem [2, Theorem 2.1]). *Let X be a real Hilbert space. Let $\beta > 0$ and suppose that $g\colon X \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function. Then the following are equivalent:*

(*i*) *g takes only real values, it is Fréchet differentiable on X and its gradient $\nabla g\colon X \to X$ is $\frac{1}{\beta}$-cocoercive[1];*

(*ii*) *$\frac{\beta}{2}\|\cdot\|^2 - g$ is convex.*

We will prove Theorem 1 by reducing it to the case $Y = \mathbb{R}$ through the following lemma.

Lemma 2. *Let X and Y be real Hilbert and Banach spaces, respectively. Let $F : X \to Y$ be a continuous function. Let $L > 0$ and suppose that (2) holds for any $x_1, \dots, x_n \in X$ and $\lambda_1, \dots, \lambda_n \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$. For each $y^* \in Y^*$, define the slice $\phi_{y^*} : X \to \mathbb{R}$ by $\phi_{y^*} := y^* \circ F$, i.e.,*

$$\phi_{y^*}(x) = y^*(F(x)) \quad \forall x \in X. \tag{3}$$

Then, for each $y^* \in Y^*$ with $\|y^*\| \leq 1$, the function ϕ_{y^*} is Fréchet differentiable everywhere and its gradient $\nabla\phi_{y^*} : X \to X$ is L-Lipschitz continuous.

Proof. Inequality (2) gives for any $x, y \in X$ and $t \in [0,1]$

$$\|F(x+t(y-x)) - (1-t)F(x) - tF(y)\| \leq \frac{L}{2}t(1-t)\|x-y\|^2. \tag{4}$$

For every $x, y \in X$ and $t \in [0,1]$, we have

[1] $G\colon X \to X$ is $\frac{1}{\beta}$-cocoercive if $\langle G(x) - G(y), x - y\rangle \geq \frac{1}{\beta}\|G(x) - G(y)\|^2$ for all $x, y \in X$.

$$
\begin{aligned}
&\left|\phi_{y^*}(x+t(y-x))-(1-t)\phi_{y^*}(x)-t\phi_{y^*}(y)\right| \\
&=\left|y^*\Big(F(x+t(y-x))-(1-t)F(x)-tF(y)\Big)\right| && \text{(by the definition (3) of } \phi_{y^*}) \\
&\le \|F(x+t(y-x))-(1-t)F(x)-tF(y)\| && \text{(by } \|y^*\|\le 1) \\
&\le \frac{L}{2}t(1-t)\|x-y\|^2, && \text{(by (4))}
\end{aligned}
$$

which implies that

$$
-\frac{L}{2}t(1-t)\|x-y\|^2 \le \phi_{y^*}(x+t(y-x))-(1-t)\phi_{y^*}(x)-t\phi_{y^*}(y) \le \frac{L}{2}t(1-t)\|x-y\|^2. \tag{5}
$$

The first and second inequalities in (5), together with the identity

$$
t(1-t)\|x-y\|^2=(1-t)\|x\|^2+t\|y\|^2-\|x+t(y-x)\|^2,
$$

imply that $\frac{L}{2}\|\cdot\|^2-\phi_{y^*}$ and $\frac{L}{2}\|\cdot\|^2+\phi_{y^*}$ are convex, respectively.

In particular, we have

$$
L\|\cdot\|^2-\underbrace{\left(\frac{L}{2}\|\cdot\|^2+\phi_{y^*}\right)}_{\text{convex}}=\underbrace{\frac{L}{2}\|\cdot\|^2-\phi_{y^*}}_{\text{convex}}.
$$

By Theorem 2, $\frac{L}{2}\|\cdot\|^2+\phi_{y^*}$ is Fréchet differentiable on X and $\nabla(\frac{L}{2}\|\cdot\|^2+\phi_{y^*})$ is $\frac{1}{2L}$-cocoercive. In particular, ϕ_{y^*} is Fréchet differentiable on X. The cocoercivity of $\nabla(\frac{L}{2}\|\cdot\|^2+\phi_{y^*})$ implies that

$$
\left\langle L(x-y)+\nabla\phi_{y^*}(x)-\nabla\phi_{y^*}(y),x-y\right\rangle \ge \frac{1}{2L}\left\|L(x-y)+\nabla\phi_{y^*}(x)-\nabla\phi_{y^*}(y)\right\|^2
$$

for all $x,y\in X$. Expanding both sides and rearranging terms yields $\|\nabla\phi_{y^*}(x)-\nabla\phi_{y^*}(y)\|^2\le L^2\|x-y\|^2$, which proves that $\nabla\phi_{y^*}$ is L-Lipschitz continuous.

The biggest hurdle in proving Theorem 1 is establishing the Fréchet differentiability of F. If Y were finite-dimensional, say, $Y=\mathbb{R}^d$, we would have $F=(F_1,\dots,F_d)$ for certain functions $F_i\colon X\to\mathbb{R}$. These F_i are, of course, the slices of F defined by the usual unit vectors of $\mathbb{R}^d$. Then, Lemma 2 would imply that all the F_i have an L-Lipschitz derivative from which we would conclude the Fréchet differentiability of F through elementary means. The case where Y is infinite-dimensional is more delicate as it takes more effort to establish that F is indeed Fréchet differentiable by analyzing its slices, as shown in the proof of the following lemma.

Lemma 3. *Let X and Y be real Banach spaces and suppose that Y is reflexive. Let $F\colon X\to Y$ be a continuous function and let $L>0$. For each $y^*\in Y^*$ with $\|y^*\|\le 1$, suppose that the slice $\phi_{y^*}\colon X\to\mathbb{R}$ defined by (3) is Fréchet differentiable and its derivative $\phi'_{y^*}\colon X\to X^*$ is L-Lipschitz continuous. Then, F is Fréchet differentiable and its derivative $F'\colon X\to\mathscr{B}(X,Y)$ is L-Lipschitz continuous.*

Note that, to make the statement more general, X is not assumed to be a Hilbert space; rather, it is only assumed to be a Banach space, unlike in Lemma 2. Therefore, we use the derivative $\phi'_{y^*} : X \to X^*$ in place of the gradient $\nabla\phi_{y^*} : X \to X$ in this lemma.

Proof (Proof of Lemma 3). By the Lipschitz continuity of ϕ'_{y^*}, for every $x, h \in X$ and $y^* \in Y^*$ with $\|y^*\| \leq 1$, we have

$$\left|\phi_{y^*}(x+h) - \phi_{y^*}(x) - \phi'_{y^*}(x)h\right| \leq \frac{L}{2}\|h\|^2. \tag{6}$$

The proof is divided into three parts:

(1) showing that for each $x \in X$, the map $y^* \mapsto \phi'_{y^*}(x)$ is a bounded linear map from Y^* to X^*,
(2) constructing the Fréchet derivative of F using the reflexivity of Y, and
(3) showing that the derivative is L-Lipschitz continuous.

(1) Linearity and boundedness of $y^* \mapsto \phi'_{y^*}(x)$ for every $x \in X$.

Fix $x \in X$ arbitrarily. The map $y^* \mapsto \phi_{y^*}(x)$ is linear by the definition (3), hence $y^* \mapsto \phi'_{y^*}(x)$ is a linear map between Y^* and X^*.

Next, we show the boundedness. The continuity of F at x implies that there exists $\delta > 0$ such that for each $h \in X$ with $\|h\| \leq \delta$ it holds

$$\|F(x+h) - F(x)\| \leq 1. \tag{7}$$

Thus, for each $y^* \in Y^*$ with $\|y^*\| \leq 1$, we have

$$\begin{aligned}
\left\|\phi'_{y^*}(x)\right\| &= \frac{1}{\delta}\sup_{\|h\|\leq\delta}\left|\phi'_{y^*}(x)h\right| \\
&\leq \frac{1}{\delta}\sup_{\|h\|\leq\delta}\left(\left|\phi_{y^*}(x+h) - \phi_{y^*}(x)\right| + \frac{L}{2}\|h\|^2\right) && \text{(by (6))} \\
&\leq \frac{1}{\delta}\sup_{\|h\|\leq\delta}\left(\|F(x+h) - F(x)\| + \frac{L}{2}\|h\|^2\right) && \text{(by the definition (3) of } \phi_{y^*} \\
&&& \quad \text{and } \|y^*\| \leq 1) \\
&\leq \frac{1}{\delta} + \frac{L}{2}\delta, && \text{(by (7))}
\end{aligned}$$

which proves the boundedness of $y^* \mapsto \phi'_{y^*}(x)$.

(2) Constructing the Fréchet derivative of F.

We have shown that for each $x \in X$, the map $y^* \mapsto \phi'_{y^*}(x)$ is a bounded linear map between Y^* and X^*. Therefore, for each $x, h \in X$, the map $y^* \mapsto \phi'_{y^*}(x)h$ is an element

of Y^{**}. The reflexivity of Y implies that for each $x,h \in X$, there exists a unique $f_x(h) \in Y$ such that

$$y^*(f_x(h)) = \phi'_{y^*}(x)h \quad \forall y^* \in Y^*. \tag{8}$$

We will show that $f_x \colon X \to Y$ (i.e., $h \mapsto f_x(h)$) is the Fréchet derivative of F at x. Let $Y_1^* \coloneqq \{\, y^* \in Y^* \mid \|y^*\| \le 1 \,\}$. The Hahn–Banach theorem implies that for any $y \in Y$ we have

$$\|y\| = \sup_{y^* \in Y_1^*} y^*(y), \tag{9}$$

see [4, Corollary 6.7]. Thus, for each $x,h \in X$, we have

$$\begin{aligned}
&\|F(x+h) - F(x) - f_x(h)\| \\
&= \sup_{y^* \in Y_1^*} y^*\Big(F(x+h) - F(x) - f_x(h)\Big) && \text{(by (9))} \\
&= \sup_{y^* \in Y_1^*} \Big(\phi_{y^*}(x+h) - \phi_{y^*}(x) - \phi'_{y^*}(x)h\Big) && \text{(by (3) and (8))} \\
&\le \frac{L}{2}\|h\|^2 = o(\|h\|). && \text{(by (6))}
\end{aligned}$$

It remains to show that for each $x \in X$ the map $h \mapsto f_x(h)$ is linear and bounded. Fix $x \in X$ arbitrarily. Eq. (8) implies that for each $h_1, h_2 \in X$ and $y^* \in Y^*$

$$y^*\Big(f_x(h_1+h_2) - f_x(h_1) - f_x(h_2)\Big) = \phi'_{y^*}(x)(h_1+h_2) - \phi'_{y^*}(x)h_1 - \phi'_{y^*}(x)h_2 = 0,$$

and hence $f_x(h_1+h_2) = f_x(h_1) + f_x(h_2)$. Similarly we have $f_x(\alpha h) = \alpha f_x(h)$ for each $\alpha \in \mathbb{R}$. Therefore, the map $h \mapsto f_x(h)$ is linear. We next show the boundedness. For each $h \in X$, we have

$$\begin{aligned}
\|f_x(h)\| &= \sup_{y^* \in Y_1^*} y^*(f_x(h)) && \text{(by (9))} \\
&= \sup_{y^* \in Y_1^*} \phi'_{y^*}(x)h && \text{(by (8))} \\
&\le \|h\| \sup_{y^* \in Y_1^*} \|\phi'_{y^*}(x)\|.
\end{aligned}$$

Since the boundedness of $y^* \mapsto \phi'_{y^*}(x)$ was shown in the first part of this proof, the map $h \mapsto f_x(h)$ is bounded as well.

(3) Lipschitz continuity of the derivative.

The Lipschitz continuity of $x \mapsto f_x$ follows from the Lipschitz continuity of ϕ'_{y^*}. Indeed, for each $x,y \in X$ it holds

$$\begin{aligned}
\big\|f_x - f_y\big\| &= \sup_{\|h\|\le 1,\ y^*\in Y_1^*} y^*(f_x(h) - f_y(h)) && \text{(by (9))}\\
&= \sup_{\|h\|\le 1,\ y^*\in Y_1^*} \big(\phi'_{y^*}(x) - \phi'_{y^*}(y)\big)h && \text{(by (8))}\\
&= \sup_{y^*\in Y_1^*} \big\|\phi'_{y^*}(x) - \phi'_{y^*}(y)\big\| \\
&\le L\|x-y\|, && \text{(by the Lipschitz continuity of } \phi'_{y^*})
\end{aligned}$$

which completes the proof.

Theorem 1 directly follows from Lemmas 2 and 3 as shown below.

Proof (Proof of (ii) $\Longrightarrow$ (i) *in Theorem 1).* Let $F\colon X \to Y$ be a continuous function satisfying (2). By Lemma 2, for each $y^* \in Y^*$ with $\|y^*\| \le 1$, the slice ϕ_{y^*} has an L-Lipschitz continuous gradient $\nabla\phi_{y^*}$. By Lemma 3, F also has an L-Lipschitz continuous derivative F'.

3 Final remarks

In the proof of (ii) $\Longrightarrow$ (i) in Theorem 1, only the case $n = 2$ of (2) is used.

As usual, it may be interesting to see if Theorem 1 can be further extended to more general settings. This would require some extension of Theorem 2, which may be nontrivial, see [9] for some discussion along these lines. Another source of difficulty is the fact that $\|\cdot\|^2$ is not ensured to be differentiable in an arbitrary normed vector space.

Acknowledgements This work was initiated during the MATRIX workshop "Splitting Algorithms – Advances, Challenges, and Opportunities". The authors would like to thank MATRIX and the organizers of the workshop for the environment that allowed this project to flourish.

This work was also partially supported by JSPS KAKENHI (24K23853) and JST CREST (JPMJCR24Q2).

References

1. Allen-Zhu, Z., Li, Y.: NEON2: Finding local minima via first-order oracles. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). URL https://papers.nips.cc/paper/by-source-2018-1873
2. Bauschke, H.H., Combettes, P.L.: The Baillon-Haddad theorem revisited. Journal of Convex Analysis **17**(3&4), 781–787 (2010)
3. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In: D. Precup, Y.W. Teh

(eds.) Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 70, pp. 654–663. PMLR (2017). URL https://proceedings.mlr.press/v70/carmon17a.html

4. Conway, J.B.: A Course in Functional Analysis. Graduate Texts in Mathematics. Springer New York (2007)
5. Jin, C., Netrapalli, P., Jordan, M.I.: Accelerated gradient descent escapes saddle points faster than gradient descent. In: S. Bubeck, V. Perchet, P. Rigollet (eds.) Proceedings of the 31st Conference On Learning Theory, *Proceedings of Machine Learning Research*, vol. 75, pp. 1042–1085. PMLR (2018)
6. Li, H., Lin, Z.: Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $O(\varepsilon^{-7/4})$ complexity. Journal of Machine Learning Research **24**(157), 1–37 (2023). URL http://jmlr.org/papers/v24/22-0522.html
7. Marumo, N., Takeda, A.: Parameter-free accelerated gradient descent for nonconvex minimization. SIAM Journal on Optimization **34**(2), 2093–2120 (2024). URL https://doi.org/10.1137/22M1540934
8. Marumo, N., Takeda, A.: Universal heavy-ball method for nonconvex optimization under Hölder continuous Hessians. Mathematical Programming **212**(1), 147–175 (2025). URL https://doi.org/10.1007/s10107-024-02100-4
9. Wachsmuth, D., Wachsmuth, G.: A simple proof of the Baillon-Haddad theorem on open subsets of Hilbert spaces. arXiv e-print (2022)
10. Xu, Y., Jin, R., Yang, T.: NEON+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. arXiv e-print (2017)