
Balancing Gradient and Hessian Queries in Non-Convex Optimization

Deeksha Adil* Brian Bullins† Aaron Sidford‡ Chenyi Zhang‡

Abstract

We develop optimization methods which offer new trade-offs between the number of gradient and Hessian computations needed to compute the critical point of a non-convex function. We provide a method that for a twice-differentiable $f: \mathbb{R}^d \rightarrow \mathbb{R}$ with L_2 -Lipschitz Hessian, an input initial point with Δ -bounded sub-optimality, and a sufficiently small $\epsilon > 0$, outputs an ϵ -critical point, i.e., a point x such that $\|\nabla f(x)\| \leq \epsilon$, using $\tilde{O}(\Delta L_2^{1/4} n_H^{-1/2} \epsilon^{-9/4})$ queries to a gradient oracle and n_H queries to a Hessian oracle. As a consequence, we obtain an improved gradient query complexity of $\tilde{O}(d^{1/3} L_2^{1/2} \Delta \epsilon^{-3/2})$ in the case of bounded dimension and of $\tilde{O}(\Delta^{3/2} L_2^{3/4} \epsilon^{-9/4})$ in the case where we are allowed only a *single* Hessian query. We obtain these results through a more general algorithm which can handle approximate Hessian computations and recovers known prior state-of-the-art bounds of computing an ϵ -critical point, under the additional assumption that f has an L_1 -Lipschitz gradient, with $O(\Delta L_2^{1/4} \epsilon^{-7/4})$ gradient queries.

1 Introduction

We consider the problem of computing an ϵ -critical point of a differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$, that is x with $\|\nabla f(x)\| \leq \epsilon$, given an initial point $x^{(0)} \in \mathbb{R}^d$ with bounded function error or sub-optimality, $\Delta := f(x^{(0)}) - \inf_{x \in \mathbb{R}^d} f(x)$.⁴ This *critical point computation problem*—also referred to as making the gradient norm small [2, 33] or finding stationary points [1]—is a foundational and well-studied optimization problem. It is ubiquitous in machine learning research and has been studied extensively for decades; see e.g., [10, 11] for references.

Obtaining an ϵ -critical point is a natural stopping condition for many optimization methods. For general smooth non-convex functions, guarantees of this type—as opposed to reaching a globally optimal point—may be provably established without incurring exponential dimension dependence in the rates [10, 11]. Furthermore, there are even instances of non-convex objectives, such as regression tasks with non-convex regularization [31] and matrix completion [21], for which reaching what is known as a second-order critical point [1, 9, 26, 27]—which generalize ϵ -critical points—suffices to establish global optimality.

In certain foundational settings, optimal query complexities for critical point computation are known. For example, consider the following simple variant of gradient descent, e.g.,

$$x^{(t+1)} \leftarrow x^{(t)} - \frac{1}{L_1} \nabla f(x^{(t)}) = \operatorname{argmin}_{x \in \mathbb{R}^d} T_{x^{(t)}}^1(x) + \frac{L_1}{2} \|x^{(t)} - x\|^2 \quad (1)$$

*Institute for Theoretical Studies, ETH Zürich deeksha.adil@eth-its.ethz.ch

†Department of Computer Science, Purdue University bullins@purdue.edu

‡Stanford University {sidford,chenyiz}@stanford.edu

Preprint available at [arXiv:2510.20786](https://arxiv.org/abs/2510.20786).

⁴Throughout the paper we let $\|\cdot\|$ denote the Euclidean or ℓ_2 norm, and when applied to a square matrix, we let it denote the ℓ_2 -operator norm, i.e., $\|A\| = \sup_{x \in \mathbb{R}^d} \|Ax\|/\|x\|$ for all $A \in \mathbb{R}^{d \times d}$.

where $T_{x^{(t)}}^p(x)$ is the p th-order Taylor approximation of f evaluated at $x^{(t)}$ and f is L_1 -smooth, i.e., has an L_1 -Lipschitz gradient where $\|\nabla f(x) - \nabla f(y)\| \leq L_1\|x - y\|$ for all $x, y \in \mathbb{R}^d$. Eq. (1) computes an ϵ -critical point with at most $O(L_1\Delta\epsilon^{-2})$ iterations [32], and it is known that for sufficiently large d , any method (even a randomized one that succeeds with high probability) must compute at least $\Omega(L_1\Delta\epsilon^{-2})$ gradients in the worst case [10]. More broadly, for f that has L_p -Lipschitz p th-order derivatives define a *p th-order oracle* as one that when queried at a point x returns all partial derivatives of f at x up to order p . It is known that the method

$$x^{(t+1)} \leftarrow \operatorname{argmin}_{x \in \mathbb{R}^d} T_{x^{(t)}}^p(x) + c_p L_p \|x^{(t)} - x\|^{p+1},$$

for suitable choice of constant $c_p > 0$, solves the problem in $O(c_p L_p^{1/p} \Delta \epsilon^{-(p+1)/p})$ queries and iterations [7], and these rates are optimal in terms of dimension-independent query complexity [10].

Unfortunately, the faster rates for critical point computation obtained via higher-order oracles, e.g., p th-order oracles, generally come at a cost. Simply specifying the output of a p th-order oracle at a point involves outputting d^p numbers in the worst case, which even for a small constant p can be prohibitively expensive when the problem dimension is large. Correspondingly, there has been extensive research [1, 9, 10, 11, 34] on what rates are obtainable given various assumptions on f and access only to a *gradient oracle*, i.e., an oracle that when queried at $x \in \mathbb{R}^d$ outputs $\nabla f(x)$, the gradient of f at x , and a *Hessian oracle*, i.e., an oracle that when queried at $x \in \mathbb{R}^d$ outputs $\nabla^2 f(x)$, the Hessian of f at x . One particularly relevant line of work has shown that with only a gradient oracle for f that has L_1 -Lipschitz gradient and L_2 -Lipschitz Hessian, it is possible to obtain a rate of $O(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$ if $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$ [1, 9, 29], and [11] provided an $\Omega(L_1^{3/7} L_2^{2/7} \Delta \epsilon^{-12/7})$ lower bound for this setting. In other words, when the Hessian of the function is also Lipschitz continuous, it is possible to improve upon the $O(L_1 \Delta \epsilon^{-2})$ query complexity of gradient descent. Meanwhile, if we further allow for querying Hessian information, this rate can be improved to $O(L_2^{1/2} \Delta \epsilon^{-3/2})$ [34], which is optimal under this stronger oracle model [10, 12].

In this paper, we perform a more fine-grained study of the problem of critical point computation. We ask, *what trade-offs are possible between the number of gradient and Hessian computations that are needed to compute ϵ -critical points for functions with L_1 -Lipschitz gradients and L_2 -Lipschitz Hessians?* Our main result is an algorithm which offers a new such trade-off, even when the Hessian is computed only approximately. Furthermore, we show that this result generalizes the bounds of [29] and yields further improvements in dimension-dependent settings.

1.1 Our Results

Our main result is a new trade-off in the number of approximate Hessian oracle queries and gradient queries for f needed for ϵ -critical point computation. Here we define the approximate Hessian oracle we consider and provide our main theorem.

Definition 1 (Approximate Hessian Oracle). *We call a procedure a δ -approximate Hessian oracle for twice differentiable $f : \mathbb{R}^d \rightarrow \mathbb{R}$ if when queried at $x \in \mathbb{R}^d$ it outputs symmetric $H_x \in \mathbb{R}^{d \times d}$ such that $\|H_x - \nabla^2 f(x)\| \leq \delta$.*

Theorem 1 (Main Result). *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ have L_1 -Lipschitz gradient and L_2 -Lipschitz Hessian. There is an algorithm which given any $x^{(0)} \in \mathbb{R}^d$ with Δ -bounded sub-optimality with respect to f , positive integer n_H , and $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, outputs an ϵ -critical point of f with at most n_H queries to a δ -approximate Hessian oracle and at most*

$$O\left(\frac{\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \operatorname{poly} \log\left(\frac{c_\ell}{c_\delta}\right)\right)$$

queries to a gradient oracle for f , where

$$c_\delta := \min\left\{L_1, \delta + \frac{\Delta L_2}{n_H \epsilon}\right\} \text{ and } c_\ell := \min\left\{L_1, \frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta^2}{\epsilon^2} + \delta\right\}.$$

As outlined in our overview in Section 2, Theorem 1 follows from a careful combination of Theorem 3 and Corollary 2, which characterize Algorithm 2 and Algorithm 3, respectively.

In the remainder of this section we compare this result to previous studied problems and discuss its implications.

Generalizing Prior Gradient Methods. First, we note that Theorem 1 recovers known prior results on gradient-only methods. Observe that $\|\nabla^2 f(x)\| \leq L_1$ if and only if f has an L_1 -Lipschitz gradient, and consequently, an L_1 -approximate Hessian oracle for f can be implemented with no queries, by simply outputting the all-zero matrix. In this case, the approximation error is $\delta = L_1$, which leads to $c_\delta = c_\ell = L_1$. Thus, as a corollary of Theorem 1, we obtain a method which computes an ϵ -critical point using $O(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$ queries, which matches the prior best known algorithms in this setting [1, 8, 9, 29].

Gradient-Hessian Trade-offs for Functions with Unbounded Smoothness. Interestingly, our results apply even without a bound on L_1 . In this case, it follows from Theorem 1 that, with exact Hessian queries, i.e., when $\delta = 0$, it is possible to obtain methods that compute an ϵ -critical point with n_H Hessian queries and $\tilde{O}(\Delta^{3/2} L_2^{3/4} n_H^{-1/2} \epsilon^{-9/4})$ gradient queries.⁵ Excitingly, this result shows it is possible to compute an ϵ -critical point with only a *single* Hessian query and $\tilde{O}(\Delta^{3/2} L_2^{3/4} \epsilon^{-9/4})$ gradient queries. The previous best algorithms in this setting are essentially due to Doikov et al. [18]. Though their paper studies a different setting, their results seem to imply an ϵ -critical point with n_H Hessian queries and $O(\Delta^2 L_2 n_H^{-2} \epsilon^{-3})$ gradient queries, and so for a single Hessian query, we improve these results, up to polylogarithmic factors, by a factor of $O(\Delta^{1/2} L_2^{1/4} \epsilon^{-3/4})$.

Dimension-Dependent Critical Point Computation. As another application of Theorem 1, we obtain improved bounds on the number of gradients needed to compute a critical point of functions where d , the dimension, is bounded. Specifically, we note that by a finite differencing argument (e.g., Lemma 3 in [23] for $h = 2\delta d^{-1/2} L_2^{-1}$), a δ -approximate Hessian oracle for f can be implemented with just $2d$ queries to a gradient oracle by approximating each column of the Hessian using by finite differences and two gradient queries. Applying this fact with Theorem 1 and optimizing over the choice of n_H yields the following corollary.

Corollary 1. *Let $f: \mathbb{R}^d \mapsto \mathbb{R}$ have L_1 -Lipschitz gradient and L_2 -Lipschitz Hessian. There is a method which given any $x^{(0)} \in \mathbb{R}^d$ with Δ -bounded sub-optimality with respect to f and $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, outputs an ϵ -critical point of f with at most $\tilde{O}(d^{1/3} L_2^{1/2} \Delta \epsilon^{-3/2} + d)$ queries to a gradient oracle for f .*

Interestingly, while optimal query complexities are known for the low-dimensional $d = 1$ [15] and $d = 2$ [24] cases—the latter following from the more general $O(\max\{2^d, \epsilon^{-2d/(d+2)}\})$ query complexity results of Vavasis [37]—for $d \geq 7$, our results improve, up to polylogarithmic factors, upon the previous state-of-the-art dimension-dependent bound of $O(\sqrt{d} L_2 \Delta \epsilon^{-3/2})$ for this problem [18] for any $\epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$. In addition, we improve upon the $O(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$ bounds of Li and Lin [29] for $\epsilon \leq \min\{O(L_1^{3/2} L_2^{-1} d^{-4/3}), L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$. This expands the range of ϵ for which the rate of [29] can be improved over [18], which offers improvement when $\epsilon \leq O(L_1^2 L_2^{-1} d^{-2})$; see Table 1 for details. Work by Jiang et al. [25] also provides dimension-dependent results of $O(d^{1/4} L_1^{1/4} L_2^{3/8} \Delta \epsilon^{-13/8})$ gradient queries under the additional assumption that $\epsilon \leq \Delta L_2 / L_1$. However, as shown in Appendix B, for any $d \geq 1$ and $\epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}, \Delta L_2 L_1^{-1}\}$, this is always at least the minimum of $O(\sqrt{d} L_2 \Delta \epsilon^{-3/2} + d)$ [18] and $O(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$ [29].

It remains unclear whether our bound, particularly the $d^{1/3}$ dependence, is asymptotically optimal. Although there are relevant lower bounds in the dimension-independent setting [10, 11], the dimension-dependent complexity of critical point computation is still not well understood: Existing dimension-dependent lower bounds apply only to low-dimensional settings and do not have Hessian Lipschitzness assumptions, see e.g., [15, 24]. The development of tight lower bounds in our regime as an independent and interesting open problem.

While not the focus of the paper, we briefly comment on the computational complexity of our algorithm. Each iteration involves an approximate eigendecomposition step, which may seem more involved than the computation in the classical Newton method, where a dense linear system is solved in each iteration. Nevertheless, in the worst case, it can be implemented in $O(d^\omega)$ time, where ω

⁵We use $\tilde{O}(\cdot)$ to hide polylogarithmic factors in n_H , $\max\{L_2, L_2^{-1}\}$, $\max\{\Delta, \Delta^{-1}\}$, $\max\{\epsilon, \epsilon^{-1}\}$, and δ .

Algorithm	# Gradient Queries	Assumption
Vavasis [37]	$O(2^d + \epsilon^{-2d/(d+2)})$	L_1
Li & Lin [29]	$O(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$	L_1, L_2
Nesterov & Polyak [34]	$O(d\sqrt{L_2} \Delta \epsilon^{-3/2})$	L_2
Doikov et al. [18]	$O(\sqrt{dL_2} \Delta \epsilon^{-3/2} + d)$	L_2
Jiang et al. [25]	$O(d^{1/4} L_1^{1/4} L_2^{3/8} \Delta \epsilon^{-13/8})^6$	L_1, L_2
Corollary 1 (Ours)	$\tilde{O}(d^{1/3} L_2^{1/2} \Delta \epsilon^{-3/2} + d)$	L_2

Table 1: Comparison with previous results in terms of the number of gradient queries needed to reach an ϵ -critical point, i.e., $\|\nabla f(x)\| \leq \epsilon$, under L_1 -Lipschitz gradient, L_2 -Lipschitz Hessian assumptions, for $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$. This is a standard range of ϵ to consider as noted in [8]. If $\epsilon > L_1^2 L_2^{-1}$, gradient descent achieves better query complexity. If $\epsilon > \Delta^{2/3} L_2^{1/3}$, our algorithm halts after at most a single iteration and makes at most $\tilde{O}(d^{1/3})$ queries. Up to polylogarithmic factors, our results improve upon [29] when $\epsilon \leq \min\{O(L_1^{3/2} L_2^{-1} d^{-4/3}), L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, and upon [18] for any $\epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$.

denotes the matrix multiplication exponent [17]. This is the same as the per-iteration cost of the Newton method.

Additional Related Work. The efficiency of critical point computation has been explored in a wide variety of non-convex optimization contexts and settings, some of which we now highlight. Although this work is concerned with exact gradient information, there has also been significant effort in understanding optimal complexities when instead given access to stochastic oracles [4, 2, 3, 5, 6, 20, 22, 40]. In addition, other works have considered methods based on alternative structural assumptions, such as a type of graded non-convexity [19], a particular spectral decay of the Hessian [30], or, in the case of *non-smooth* non-convex objectives, relaxed notions of approximate critical points [16, 28, 36, 39]. There is also a broader literature related to general Taylor descent algorithms [7, 10], including works that focus on efficient and adaptive methods [13, 14, 34].

2 Our Algorithms

Our approach is inspired by the advances in obtaining $\tilde{O}(L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4})$ rates over the last few years [1, 9, 29]. In particular, our algorithm builds on the work of [29], which proves that one can apply accelerated gradient descent with restarts to obtain improved rates. Our algorithm uses a similar method, but works with a norm induced by computations of the approximate Hessian H . In particular, we work in the norm induced by $\phi(H)$ where ϕ is a carefully chosen function which returns a symmetric matrix. We show that by applying their method in this carefully designed norm and recomputing the Hessian intermittently, we obtain our result.

First, in Section 2.1 we present a variant of accelerated gradient descent (Algorithm 1). This is similar to the algorithms of [29] without restarts, but in the norm induced by $\hat{H} := \phi(H)$ (Eq. 2). We prove that with a single Hessian computation and a bounded number of gradient computations, the algorithm either finds a critical point or significantly reduces the function value (Theorem 2).

Second, to make use of this result, we either perform negative curvature descent whenever the approximate Hessian H has a sufficiently negative eigenvalue, or apply a restart strategy similar to [29]. (Algorithm 2). Our algorithm additionally keeps track of the movement of the iterates and when the movement is too large, recomputes the approximate Hessian H and the corresponding $\hat{H}$. In Section 2.2 we analyze Algorithm 2 which essentially obtains our main result up to logarithmic factors (Theorem 3).

Unfortunately, the logarithmic factors for Algorithm 2 depend on L_1 . Interestingly, we show that there is a fairly generic procedure that allows us to remove this dependence with at most one additional Hessian computation. In Section 2.3, we prove a general reduction that given an algorithm that finds critical points for a function with L_1 -Lipschitz gradient and L_2 -Lipschitz Hessian, there is an

⁶This result holds under the additional assumption that $\epsilon = O(\Delta L_2 / L_1)$. See Appendix B for a detailed comparison between this work and [18, 29].

algorithm that uses one additional Hessian computation and finds a critical point for any function with only an L_2 -Lipschitz Hessian (Theorem 4). Finally, with all of these tools in hand, we prove our main result, Theorem 1. Several proofs are deferred to the appendix.

2.1 Critical or Progress using Approximate Hessians

In this section, we describe the core subroutine of our critical point computation algorithm, **Critical-or-Progress**, which is a version of accelerated gradient descent (Algorithm 1). This procedure either finds an ϵ -critical point or decreases the function value by at least $\tilde{\Omega}(\epsilon^{3/2}/L_2^{1/2})$. Suppose that we are given $H_{x^{(0)}}$, a δ -approximate Hessian at $x^{(0)}$, such that $\|H_{x^{(0)}} - \nabla^2 f(x^{(0)})\| \leq \delta$, and additionally $-2\delta I \preceq H_{x^{(0)}} \preceq L_1 I$. Let the spectral decomposition of $H_{x^{(0)}}$ be

$$H_{x^{(0)}} = \sum_{j=1}^d \lambda_j h_j h_j^\top,$$

where $\{h_1, \dots, h_d\}$ is an orthonormal basis. Define

$$\hat{H} := \phi(\lambda_j) h_j h_j^\top, \quad (2)$$

where

$$\phi(\lambda) := (32\delta + |\lambda|) \cdot \frac{\lceil \log_2(L_1/\delta) \rceil}{\lceil \log_2(\max\{|\lambda|, 2\delta\}/\delta) \rceil}. \quad (3)$$

We consider an algorithm that performs AGD in the norm induced by $\hat{H}$, as shown in Algorithm 1. The parameters used in Algorithm 1 are chosen as follows.

$$\begin{aligned} p_{\max} &= \max\{\lceil \log(L_1/\delta) \rceil, 16\}, \quad \tilde{\epsilon} = \frac{\epsilon}{p_{\max}^8}, \quad \eta = \frac{1}{4}, \\ B &= \frac{1}{3} \sqrt{\frac{\tilde{\epsilon}}{L_2}}, \quad \theta = \frac{1}{K}, \quad K = \frac{\sqrt{\delta}}{(\tilde{\epsilon} L_2)^{1/4}} \end{aligned} \quad (4)$$

Algorithm 1: Critical-or-Progress

- 1 **Input:** initial iterate $x^{(0)}$, a δ -approximate Hessian $H_{x^{(0)}}$, target accuracy ϵ , gradient Lipschitzness L_1 , Hessian Lipschitzness L_2 ;
 - 2 Initialize $x^{(-1)} \leftarrow x^{(0)}$;
 - 3 Set $\hat{H}$ according to (2), and set η, B, θ, K according to (4);
 - 4 **for** $k = 0, \dots, K$ **do**
 - 5 $y^{(k)} \leftarrow x^{(k)} + (1 - \theta_1)(x^{(k)} - x^{(k-1)})$;
 - 6 $x^{(k+1)} \leftarrow y^{(k)} - \eta \hat{H}^{-1} \nabla f(y^{(k)})$;
 - 7 **if** $k \sum_{\kappa=0}^{k-1} \|\hat{H}^{1/2}(x^{(\kappa+1)} - x^{(\kappa)})\|^2 \geq 12\delta p_{\max} B^2$ **then Output** $x^{\text{out}} \leftarrow x^{(k)}$;
 - 8 $K_0 \leftarrow \operatorname{argmin}_{\lfloor \frac{3K}{4} \rfloor \leq k \leq K-1} \|\hat{H}^{1/2}(x^{(k+1)} - x^{(k)})\|$;
 - 9 **Output** $x^{\text{out}} \leftarrow \frac{1}{K_0+1-\lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} y^{(k)}$
-

The main result of this section is the following Theorem 2, which shows that Algorithm 1, using a single query to a δ -approximate Hessian oracle and a bounded number of gradient queries for f , either finds an ϵ -critical point or decreases the function value by at least $\tilde{\Omega}(\epsilon^{3/2}/L_2^{1/2})$.

Theorem 2. *Let $\delta \leq L_1$, $\epsilon \leq \delta^2/L_2$, and $-\delta I \preceq H_{x^{(0)}} \preceq 2L_1 I$. Using the parameters in (4), Algorithm 1 makes 1 query to a δ -approximate Hessian oracle and at most K queries to a gradient oracle for f and outputs $x^{\text{out}} \in \mathbb{R}^d$ with $\|x^{\text{out}} - x^{(0)}\| \leq 7B$ such that either x^{out} is ϵ -critical for f or*

$$f(x^{\text{out}}) - f(x^{(0)}) \leq -L_2^{-1/2} \tilde{\epsilon}^{3/2} = -L_2^{-1/2} p_{\max}^{-12} \epsilon^{3/2}.$$

For any $x, y \in \mathbb{R}^d$, define the variables $\hat{x} := \hat{H}^{1/2}x$, $\hat{y} := \hat{H}^{1/2}y$, and the function in the norm induced by $\hat{H}$,

$$\hat{f}(\hat{x}) := f(\hat{H}^{-1/2}\hat{x}), \quad (5)$$

which satisfies

$$\nabla \hat{f}(\hat{x}) = \hat{H}^{-1/2} \cdot \nabla f(\hat{H}^{-1/2}\hat{x}). \quad (6)$$

In this norm induced by $\hat{H}$, the updates in Line 5 and Line 6 of Algorithm 1 become

$$\begin{aligned} \hat{y}^{(k)} &\leftarrow \hat{x}^{(k)} + (1 - \theta)(\hat{x}^{(k)} - \hat{x}^{(k-1)}), \\ \hat{x}^{(k+1)} &\leftarrow \hat{y}^{(k)} - \eta \nabla \hat{f}(\hat{y}^{(k)}), \end{aligned} \quad (7)$$

which are similar to the standard accelerated gradient descent updates.

The proof of Theorem 2 proceeds by analyzing whether the “if condition” in Line 7 is triggered. In the case where the iterates move relatively far in K iterations and the “if condition” is triggered, using a similar proof strategy as [29], we demonstrate that the function value must have decreased by at least $\tilde{\Omega}(L_2^{-1/2}\epsilon^{3/2})$. On the other hand, in the case where the iterates stay relatively close to $x^{(0)}$ and the “if condition” is not triggered, we show that averaging over several iterates yields a point with a small gradient. This part of the analysis is more intricate and relies sensitively on the choice of the matrix $\hat{H}$. We first establish that $\nabla \hat{f}(\hat{x}^{\text{out}})$, the gradient of the output measured in the norm induced by $\hat{H}$, is small. However, this does not immediately imply that $\nabla f(x^{\text{out}})$ is small. To bridge this gap, we leverage the specific structure of $\hat{H}$ defined in (2), where spectral gaps are intentionally introduced through the use of a piecewise function defined in its construction. Using results from matrix perturbation theory, we show that this construction guarantees that the eigenvectors of $\nabla^2 \hat{f}(\hat{x}^{(0)})$ with large eigenvalues is nearly identical to that of $\hat{H}$ and $\nabla^2 f(x^{(0)})$. Consequently, we can bound the component of $\nabla f(x^{\text{out}})$ in the strongly convex subspace using the corresponding component of $\nabla \hat{f}(\hat{x}^{\text{out}})$, and analyze the component in the non-strongly convex subspace similarly.

2.2 Restarted Approximate Hessian AGD

In this section, we present our main algorithm, Algorithm 2. In each iteration, Algorithm 2 maintains a Hessian estimate of $\nabla^2 f(x^{(t)})$ with bounded error. If the estimate has a negative eigenvalue smaller than $-3\tilde{\delta}$, we identify a direction of negative curvature and update along it. Otherwise, we invoke **Critical-or-Progress** (Algorithm 1) and set its output as the next iterate. We show that if $x^{(t+1)}$ is not an ϵ -critical point, the function value decreases efficiently. As a result, the algorithm finds an ϵ -critical point in a bounded number of iterations.

The Hessian estimate in Algorithm 2 is maintained via lazy updates. It initializes with $x^{(0)}$ as the reference point $\bar{x}$ and sets the Hessian estimate to $H_{x^{(0)}}$. Whenever the current iterate $x^{(t)}$ moves more than a threshold R away from $\bar{x}$, we update the reference point $\bar{x} \leftarrow x^{(t)}$ and the Hessian estimate to $H_{x^{(t)}}$. Since f has an L_2 -Lipschitz Hessian, the error of the Hessian estimate is bounded by $\max\{2L_1, \delta + L_2R\}$ in each iteration.

The parameters in Algorithm 2 are chosen as follows.

$$\begin{aligned} R &\leftarrow \frac{3\Delta}{n_H\epsilon} \log^8 \left(\frac{L_1}{\delta + 3L_2\Delta/(n_H\epsilon)} + 16 \right), \\ \tilde{\delta} &\leftarrow \min\{\delta + L_2R, 2L_1\}, \\ \tilde{p} &\leftarrow \max\{\lceil \log(L_1/\tilde{\delta}) \rceil, 16\}. \end{aligned} \quad (8)$$

The following theorem shows that Algorithm 2 outputs an ϵ -critical point using a bounded number of queries to a δ -approximate Hessian oracle and a gradient oracle.

Theorem 3. *Let $f: \mathbb{R}^d \mapsto \mathbb{R}$ have L_1 -Lipschitz gradients and L_2 -Lipschitz Hessian. For any $x^{(0)} \in \mathbb{R}^d$ with Δ -bounded sub-optimality with respect to f , positive integer n_H , and $0 < \epsilon \leq$*

Algorithm 2: Restarted-Approx-Hessian-AGD

1 **Input:** initial iterate $x^{(0)}$, accuracy of the approximate Hessian oracle δ , target accuracy ϵ ,
gradient Lipschitzness L_1 , Hessian Lipschitzness L_2 ;
2 Initialize $\bar{x} \leftarrow x^{(0)}$, $H \leftarrow H_{\bar{x}}$;
3 **for** $t = 0, 1, 2, \dots$ **do**
4 **if** $\|\nabla f(x^{(t)})\| \leq \epsilon$ **then** **Output** $x^{(t)}$;
5 **if** $\|x^{(t)} - \bar{x}\| \geq R$ **then** $\bar{x} \leftarrow x^{(t)}$ and $H \leftarrow H_{\bar{x}}$;
6 **if** $H \prec -3\tilde{\delta}I$ **then**
7 Choose a unit vector $v \in \mathbb{R}^d$ such that $v^\top H v \leq -2\tilde{\delta}$ and $\langle v, \nabla f(x^{(t)}) \rangle \leq 0$;
8 $x^{(t+1)} \leftarrow x^{(t)} + Rv$
9 **else** $x^{(t+1)} \leftarrow \text{Critical-or-Progress}(x^{(t)}, H, 4\tilde{\delta}, \epsilon, L_1, L_2)$;

$\min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, Algorithm 2 outputs an ϵ -critical point with at most n_H queries to a δ -approximate Hessian oracle and at most

$$\frac{2\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \log^{18} \left(\frac{L_1}{c_\delta} + 16 \right)$$

queries to a gradient oracle for f , where

$$c_\delta := \min \left\{ L_1, \delta + \frac{\Delta L_2}{n_H \epsilon} \right\}.$$

The following describes the change in function value in every iteration of Algorithm 2 and is useful for proving Theorem 3.

Lemma 1. Suppose $\epsilon \leq L_1^2/L_2$. In each iteration t of Algorithm 2 before it terminates, we have

$$f(x^{(t+1)}) - f(x^{(t)}) \leq -\tilde{p}^{-12} \sqrt{\epsilon^3/L_2}$$

if $x^{(t+1)}$ is not ϵ -critical for f , where we denote $\tilde{p} = \max\{\lceil \log(L_1/\tilde{\delta}) \rceil, 16\}$.

Proof of Theorem 3. By Lemma 1, before Algorithm 2 terminates, in each iteration the function value decreases by at least $\tilde{p}^{-12} \sqrt{\epsilon^3/L_2}$. Hence, Algorithm 2 terminates in at most $\tilde{p}^{12} \Delta \sqrt{L_2/\epsilon^3}$ iterations. From Theorem 2, the number of gradient queries in each iteration is at most

$$K = \frac{\tilde{p}^2 \sqrt{\tilde{\delta}}}{(L_2 \epsilon)^{1/4}}.$$

Therefore, the total number of gradient queries is at most

$$\begin{aligned} \tilde{p}^{12} \Delta \sqrt{\frac{L_2}{\epsilon^3}} \cdot \frac{\tilde{p}^2 \sqrt{\tilde{\delta}}}{(L_2 \epsilon)^{1/4}} &= \frac{2\tilde{p}^{18} \Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\min \left\{ L_1, \delta + \frac{\Delta L_2}{n_H \epsilon} \right\}} \\ &= \frac{2\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\min \left\{ L_1, \delta + \frac{\Delta L_2}{n_H \epsilon} \right\}} \cdot \log^{18} \left(\frac{L_1}{\tilde{\delta}} + 16 \right) \\ &\leq \frac{2\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\min \left\{ L_1, \delta + \frac{\Delta L_2}{n_H \epsilon} \right\}} \cdot \log^{18} \left(\frac{L_1}{c_\delta} + 16 \right) \end{aligned}$$

given that $c_\delta < \tilde{\delta}$. Furthermore, from Theorem 2 and (4), in every iteration t we have

$$\|x^{(t+1)} - x^{(t)}\| \leq \frac{7}{3\tilde{p}^4} \sqrt{\frac{\epsilon}{L_2}}.$$

Therefore, the number of Hessian computations is at most

$$\left\lceil \tilde{p}^{12} \Delta \sqrt{\frac{L_2}{\epsilon^3}} \cdot \frac{7}{3\tilde{p}^4} \sqrt{\frac{\epsilon}{L_2}} \cdot \frac{1}{R} \right\rceil,$$

which is at most n_H when using the values of the parameters in (8). $\square$

2.3 Removing the L_1 -Lipschitz Gradient Assumption

In this section, we give an algorithm that computes an ϵ -critical point for functions with L_2 -Lipschitz Hessians but no guarantees on the Lipschitzness of the gradient. In order to do so, we first provide a general reduction in Algorithm 3. Given an algorithm Alg for computing an approximate critical point for a function with L_1 -Lipschitz gradient and L_2 -Lipschitz Hessian, Algorithm 3 can compute an approximate critical point for any function with only L_2 -Lipschitz Hessian while maintaining the number of queries to gradient oracle and using at most one more query to the Hessian oracle, if the output of Alg is not very far away from the initial iterate and has bounded suboptimality.

For any symmetric $M \in \mathbb{R}^{d \times d}$ with spectral decomposition $M = \sum_{j \in [d]} \lambda_j v_j v_j^\top$, we denote $M^\dagger := \sum_{j \in [d]} \mathbb{I}\{\lambda_j \neq 0\} \cdot \lambda_j^{-1} v_j v_j^\top$ to be its Moore–Penrose pseudoinverse, where $\mathbb{I}\{\cdot\}$ is the indicator function.

Algorithm 3: Reduction-To-Unbounded-Hessian

- 1 **Input:** initial iterate $x^{(0)}$, target accuracy ϵ , Hessian Lipschitzness L_2 ;
 - 2 Choose an eigenvalue threshold ℓ ;
 - 3 Initialize $H \leftarrow H_{x^{(0)}}$;
 - 4 $x^{\text{out}} \leftarrow \text{Alg}(f_{\leq \ell}, \ell, L_2, \delta, \Delta, \epsilon)$;
 - 5 **Output** $y \leftarrow x^{\text{out}} - (\Pi_{>\ell} H \Pi_{>\ell})^\dagger \nabla f(x^{\text{out}})$
-

Given a δ -approximate Hessian $H_{x^{(0)}}$ with spectral decomposition $H_{x^{(0)}} = \sum_{j=1}^d \lambda_j h_j h_j^\top$, where $\{h_1, \dots, h_d\}$ is an orthonormal basis, we partition the indices into two sets based on the eigenvalue threshold ℓ picked in Line 2.

$$\mathcal{S}_{\leq \ell} := \{i : |\lambda_i| \leq \ell\}, \quad \mathcal{S}_{>\ell} := \{i : |\lambda_i| > \ell\}, \quad (9)$$

and define corresponding projection matrices

$$\Pi_{\leq \ell} := \sum_{i \in \mathcal{S}_{>\ell}} h_i h_i^\top, \quad \Pi_{>\ell} := \sum_{i \in \mathcal{S}_{\leq \ell}} h_i h_i^\top. \quad (10)$$

Moreover, we define the restricted function

$$f_{\leq \ell} := f(x^{(0)} + \Pi_{\leq \ell}(x - x^{(0)})). \quad (11)$$

Algorithm 3 begins by invoking the subroutine Alg to find an $\epsilon/2$ -stationary point x^{out} of the restricted function $f_{\leq \ell}$. It then performs a Newton step in the subspace spanned by $h_j : j \in \mathcal{S}_{>\ell}$:

$$y \leftarrow x^{\text{out}} - (\Pi_{>\ell} H_{x^{(0)}} \Pi_{>\ell})^\dagger \nabla f(x^{\text{out}})$$

using $H_{x^{(0)}}$ as the approximation of $\nabla^2 f(x^{\text{out}})$. Since $\|x^{\text{out}} - x^{(0)}\| \leq R_{\text{out}}$, we show that $H_{x^{(0)}}$ remains a sufficiently accurate estimate. Moreover, ℓ is chosen large enough so that the size of this Newton step is small. Otherwise it would incur a function value decrease larger than Δ_{out} , contradicting to the suboptimality condition of x^{out} . As a result, $\nabla f(y)$ is close to $\nabla T_{x^{\text{out}}}^2(y)$ since f is L_2 -Hessian Lipschitz, and the latter is at most $\epsilon/2$. Formally, we prove the following:

Theorem 4. Let $\text{Alg}(f_{\leq L_1}, L_1, L_2, \delta, \Delta, \epsilon)$ be a procedure that, for any function $f_{\leq L_1} : \mathbb{R}^d \rightarrow \mathbb{R}$ with L_1 -Lipschitz gradient, L_2 -Lipschitz Hessian, and Δ -bounded suboptimality, uses

- n_H queries to a δ -approximate Hessian oracle for $f_{\leq L_1}$, and,
- $n_g(L_1, L_2, \delta, \Delta, \epsilon)$ queries to a gradient oracle for $f_{\leq L_1}$,

and returns an $\epsilon/2$ -critical point x^{out} satisfying $\|x^{\text{out}} - x^{(0)}\| \leq R_{\text{out}}$ and $f(x^{\text{out}}) - \inf_z f(z) \leq \Delta_{\text{out}}$. Then, for any f with L_2 -Lipschitz Hessian and Δ -bounded suboptimality, any $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, and any ℓ that satisfies

$$\ell \geq \max \left\{ \frac{800\Delta}{\epsilon^2} (L_2 R_{\text{out}} + \delta)^2, \frac{48L_2 \Delta_{\text{out}}}{\epsilon}, 24\Delta_{\text{out}}^{1/3} L_2^{2/3}, 2\delta \right\}. \quad (12)$$

Algorithm 3 returns an ϵ -critical point using

- $n_H + 1$ queries to a δ -approximate Hessian oracle for f , and,
- $n_g(\ell, L_2, \delta, \Delta, \epsilon)$ queries to a gradient oracle for f .

The proof of Theorem 4 is in Appendix G. Corollary 2 is obtained by running Algorithm 3 where we use Algorithm 2 as the subroutine Alg. The proof of Corollary 2 is in Appendix F.

Corollary 2. *Let $f: \mathbb{R}^d \mapsto \mathbb{R}$ L_2 -Lipschitz Hessian. Given any $x^{(0)} \in \mathbb{R}^d$ with Δ -bounded sub-optimality with respect to f , any positive integer $n_H \geq 1$, and $0 < \epsilon \leq \Delta^{2/3} L_2^{1/3}$, Algorithm 3 using Algorithm 2 as the subroutine Alg outputs an ϵ -critical point of f with at most n_H queries to a δ -approximate Hessian oracle and*

$$O\left(\frac{\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\delta + \frac{\Delta L_2}{n_H \epsilon}} \cdot \text{poly log} \left(\frac{1}{c_\delta} \left(\frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta^2}{\epsilon^2} + \delta \right) \right)\right)$$

queries to a gradient oracle for f .

2.4 Proof of Theorem 1

Proof. We consider the following algorithm. When

$$L_1 \leq \frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta}{\epsilon^2} + \delta,$$

we run Algorithm 2, which outputs an ϵ -critical point with at most n_H queries to a δ -approximate Hessian oracle and at most

$$\frac{2\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \log^{18} \left(\frac{L_1}{c_\delta} + 16 \right) = O\left(\frac{\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \text{poly log} \left(\frac{c_\ell}{c_\delta} \right)\right).$$

Otherwise, we run Algorithm 3 using Algorithm 2 as the subroutine Alg, which outputs an ϵ -critical point of f with at most n_H queries to a δ -approximate Hessian oracle and

$$O\left(\frac{\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\delta + \frac{\Delta L_2}{n_H \epsilon}} \cdot \text{poly log} \left(\frac{c_\ell}{c_\delta} \right)\right) = O\left(\frac{\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \text{poly log} \left(\frac{c_\ell}{c_\delta} \right)\right)$$

queries to a gradient oracle for f , where the last equality follows from the fact that

$$L_1 > \frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta}{\epsilon^2} + \delta \geq \delta + \frac{\Delta L_2}{n_H \epsilon}$$

and thus $c_\delta = \delta + \frac{\Delta L_2}{n_H \epsilon}$. We conclude by noticing that the number of Hessian queries in both cases is n_H , while the number of gradient queries in both cases is

$$O\left(\frac{\Delta L_2^{1/4} c_\delta^{1/2}}{\epsilon^{7/4}} \cdot \text{poly log} \left(\frac{c_\ell}{c_\delta} \right)\right).$$

□

3 Conclusion

In this paper we provided new algorithms for computing critical points of twice differentiable functions using gradient and δ -approximate Hessian queries. We provided a general result which offered new trade-offs between the number of queries made to these oracles to compute an ϵ -critical point of functions with L_1 -Lipschitz gradients and L_2 -Lipschitz Hessians given an initial point of bounded suboptimality. As a consequence of this result, for sufficiently small ϵ , we recovered known bounds on the number gradient queries needed to compute critical points and improved upon the prior state-of-the-art bounds in the case where the function is either of bounded dimension or when a single Hessian query is available.

Though our work provides new algorithms and tools for critical point computation, there are several limitations to the result. First, this work is primarily theoretical, no practical implementation or

experiments are provided, and in certain cases our bounds incur multiple logarithmic factors. Second, many functions in practice may be non-differentiable or of a large enough size that computing the Hessian is prohibitively expensive, limiting the direct applicability of the results. Third, though there are interesting relevant lower bounds [10, 11], it is unknown whether our query complexities are asymptotically optimal. Each of these limitations suggests natural open problems and directions for future work, e.g., finding practical applications of our techniques and seeking improved upper and lower bounds for the problems we consider. However, we hope this paper provides valuable tools for this potential future work.

Acknowledgments

Thank you to anonymous reviewers for their feedback. Deeksha Adil is supported by Dr. Max Rössler, the Walter Haefner Foundation and the ETH Zürich Foundation. Aaron Sidford was supported in part by NSF Grant CCF-1955039. Chenyi Zhang was supported in part by the Shoucheng Zhang graduate fellowship.

References

- [1] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.
- [2] Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. *Advances in Neural Information Processing Systems*, 31, 2018.
- [3] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. In *Advances in Neural Information Processing Systems*, pages 2675–2686, 2018.
- [4] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pages 3716–3726, 2018.
- [5] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.
- [6] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.
- [7] Ernesto G. Birgin, J. L. Gardenghi, José Mario Martínez, Sandra Augusta Santos, and Ph. L. Toint. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Mathematical Programming*, 163(1):359–368, 2017.
- [8] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. “Convex until proven guilty”: Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pages 654–663. PMLR, 2017.
- [9] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.
- [10] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1):71–120, 2020.
- [11] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.
- [12] Coralia Cartis, Nicholas I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.
- [13] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

- [14] Coralía Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part II: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011.
- [15] Sinho Chewi, Sébastien Bubeck, and Adil Salim. On the complexity of finding stationary points of smooth functions in one dimension. In *International Conference on Algorithmic Learning Theory*, pages 358–374. PMLR, 2023.
- [16] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanhao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in Neural Information Processing Systems*, 35:6692–6703, 2022.
- [17] James Demmel, Ioana Dumitriu, and Olga Holtz. Fast linear algebra is stable. *Numerische Mathematik*, 108(1):59–91, 2007.
- [18] Nikita Doikov, El Mahdi Chayti, and Martin Jaggi. Second-order optimization with lazy Hessians. In *International Conference on Machine Learning*, pages 8138–8161. PMLR, 2023.
- [19] Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. Spectral preconditioning for gradient methods on graded non-convex functions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 11227–11252, 2024.
- [20] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.
- [21] Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2981–2989, 2016.
- [22] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- [23] Geovani Nunes Grapiglia, Max L. N. Gonçalves, and G. N. Silva. A cubic regularization of Newton’s method with finite difference Hessian approximations. *Numerical Algorithms*, pages 1–24, 2022.
- [24] Alexandros Hollender and Emmanouil Zampetakis. The computational complexity of finding stationary points in non-convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5571–5572. PMLR, 2023.
- [25] Ruichen Jiang, Aryan Mokhtari, and Francisco Patitucci. Improved complexity for smooth nonconvex optimization: A two-level online learning approach with quasi-Newton methods. *arXiv preprint arXiv:2412.02175*, 2025.
- [26] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1724–1732, 2017.
- [27] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory*, pages 1042–1085, 2018.
- [28] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.
- [29] Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the in the $O(\epsilon^{-7/4})$ complexity. *Journal of Machine Learning Research*, 24(157):1–37, 2023.
- [30] Yuanshi Liu, Hanzhen Zhao, Yang Xu, Pengyun Yue, and Cong Fang. Accelerated gradient algorithms with adaptive subspace search for instance-faster optimization. *arXiv preprint arXiv:2312.03218*, 2023.

- [31] Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.
- [32] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.
- [33] Yurii Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, (88):10–11, 2012.
- [34] Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
- [35] Gilbert W. Stewart and Ji-guang Sun. *Matrix perturbation theory*. 1990.
- [36] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.
- [37] Stephen A. Vavasis. Black-box complexity of local minimization. *SIAM Journal on Optimization*, 3(1):60–80, 1993.
- [38] Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.
- [39] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.
- [40] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of Machine Learning Research*, 21(103):1–63, 2020.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Refer to the Introduction section.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Refer to the Conclusions section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Refer to the Theorem statements and the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: There are no experiments in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: There are no experiments in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: There are no experiments in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: There are no experiments in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: There are no experiments in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The paper builds on theoretical foundations and has no direct societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Overview and Notation of the Appendix

The appendix is organized as follows. Appendix B provides a comparison between [25] and prior works. Appendix C presents key properties of the matrix $\widehat{H}$ defined in (2). Appendix D collects useful results from matrix perturbation theory, which are used to analyze the spectral properties of the matrices appearing in Algorithm 1. Then, we give the analysis of Algorithm 1 and prove Theorem 2 in Appendix E. The analyses of Algorithm 2 and Algorithm 3 are given in Appendix F and Appendix G, respectively.

Notation. For any symmetric matrix $M \in \mathbb{R}^{d \times d}$ with spectral decomposition $M = \sum_{j \in [d]} \lambda_j v_j v_j^\top$, and any subset $\mathcal{S} \subseteq \mathbb{R}$, we denote

$$\Pi_{\mathcal{S}}(M) := \sum_{j \in [d]} \mathbb{I}\{\lambda_j \in \mathcal{S}\} v_j v_j^\top$$

to be the projector onto the eigenspace of M with eigenvalues in $\mathcal{S}$. For any matrix $M \in \mathbb{R}^{d_1 \times d_2}$, we use $\lambda_{\min}(M)$ to denote its smallest eigenvalue. For any two invertible symmetric matrices $M, N \in \mathbb{R}^{d \times d}$ that commute, i.e., $MN = NM$, denote $\frac{M}{N} := MN^{-1} = N^{-1}M$. Moreover, we define

$$p_{\max} := \max\{\lceil \log_2(L_1/\delta) \rceil, 16\}, \quad p_{\min} := 16.$$

as in (4), where L_1 is the gradient Lipschitzness of f . For any positive integer $p \in \mathbb{N}^+$ we define

$$l_p := \frac{1}{p_{\max}} \cdot \frac{p+1}{1+2^{-(p-5)}}, \quad r_p := \frac{1}{p_{\max}} \cdot \frac{p+1}{1+2^{-(p-4)}}. \quad (13)$$

and

$$\xi_p := \frac{\sqrt{p}}{2^{p/2} p_{\max}}, \quad \bar{l}_p := l_p - \xi_p, \quad \forall p \in \mathbb{N}^+. \quad (14)$$

B Comparison Between [25] and Prior Works

In this section, we provide a comparison between [25] and prior works [18] and [29]. In particular, [25] achieves a query complexity of $O(d^{1/4} L_1^{1/4} L_2^{3/8} \Delta \epsilon^{-13/8})$, with an implicit requirement that $\epsilon \leq O(\Delta L_2/L_1)$. This condition arises from the fact that, in the first displayed equation of [25, Section C.3], the third term in the bracket needs to dominate the first two. Substituting the choices of D and M specified in [25, Theorem 4.1] yields the corresponding inequality. In the following lemma, we show that within the parameter regime $\epsilon \leq \min\{L_1^2/L_2, \Delta^{2/3} L_2^{1/3}, \Delta L_2/L_1\}$, the query complexity of [25] is asymptotically at least the minimum of [18] and [29].

Lemma 2. For $d, L_1, L_2, \Delta, \epsilon > 0$ satisfying

$$\epsilon \leq \min\left\{L_1^2/L_2, \Delta^{2/3} L_2^{1/3}, \Delta L_2/L_1\right\},$$

we have

$$\min\{\sqrt{d L_2} \Delta \epsilon^{-3/2} + d, L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4}\} = O(d^{1/4} L_1^{1/4} L_2^{3/8} \Delta \epsilon^{-13/8}).$$

Proof. Define

$$A := \sqrt{d L_2} \Delta \epsilon^{-3/2}, \quad B := L_1^{1/2} L_2^{1/4} \Delta \epsilon^{-7/4}, \quad G := d^{1/4} L_1^{1/4} L_2^{3/8} \Delta \epsilon^{-13/8}$$

and let $\phi = B/A$ and $u = d/A$. Since $G = \sqrt{AB}$,

$$\frac{\min\{A + d, B\}}{G} = \min\left\{\frac{1+u}{\sqrt{\phi}}, \sqrt{\phi}\right\}.$$

where

$$\phi = \frac{L_1^{1/2}}{\sqrt{d L_2}} \epsilon^{-1/4}, \quad u = \frac{\sqrt{d}}{\sqrt{L_2} \Delta} \epsilon^{3/2} = \frac{L_1^{1/2}}{L_2^{3/4} \Delta} \frac{\epsilon^{5/4}}{\phi}.$$

Using the condition that $\epsilon \leq \min \left\{ L_1^2/L_2, \Delta^{2/3}L_2^{1/3}, \Delta L_2/L_1 \right\}$, we obtain

$$u \leq \frac{1}{\phi} \min\{C_1, C_2, C_3\}, \quad C_1 = \frac{L_1^3}{L_2^2\Delta}, \quad C_2 = L_1^{1/2}\Delta^{-1/6}L_2^{-1/3}, \quad C_3 = \Delta^{1/4}L_2^{1/2}L_1^{-3/4}.$$

When $L_1 = \Delta^{1/3}L_2^{2/3}$, we have $C_1 = C_2 = C_3 = 1$. Moreover, since C_1 and C_2 increase with L_1 while C_3 decreases,

$$\min\{C_1, C_2, C_3\} \leq 1$$

for all L_1, L_2, Δ , and thus $u \leq 1/\phi$. Hence,

$$\frac{\min\{A+d, B\}}{G} \leq \min_{\phi \in \mathbb{R}} \left\{ \frac{1+1/\phi}{\sqrt{\phi}}, \sqrt{\phi} \right\} \leq \sqrt{\frac{1+\sqrt{5}}{2}},$$

which gives

$$\min\{\sqrt{dL_2}\Delta\epsilon^{-3/2} + d, L_1^{1/2}L_2^{1/4}\Delta\epsilon^{-7/4}\} = O(d^{1/4}L_1^{1/4}L_2^{3/8}\Delta\epsilon^{-13/8}).$$

□

C Properties of $\hat{H}$

In this section, we present several properties of the matrix $\hat{H}$ used in Algorithm 1. Let us recall the setting of Section 2.1: We are given $H_{x^{(0)}}$, a δ -approximate Hessian at $x^{(0)}$ that satisfies $\|H_{x^{(0)}} - \nabla^2 f(x^{(0)})\| \leq \delta$ and $-3\delta I \preceq H_{x^{(0)}} \preceq L_1 I$. Let the spectral decomposition be

$$H_{x^{(0)}} = \sum_{j=1}^d \lambda_j h_j h_j^\top,$$

where $\{h_1, \dots, h_d\}$ is an orthonormal basis. Then, $\hat{H}$ is defined as

$$\hat{H} := \phi(\lambda_j) h_j h_j^\top,$$

as in (2), where

$$\phi(\lambda) := (32\delta + |\lambda|) \cdot \frac{\lceil \log_2(L_1/\delta) \rceil}{\lceil \log_2(\max\{|\lambda|, 2\delta\}/\delta) \rceil}.$$

as in (3).

Lemma 3. For any positive integer $p \in \mathbb{N}^+$ and any λ such that $2^p \delta < |\lambda| \leq 2^{p+1} \delta$, we have

$$l_p < |\lambda| \cdot \phi(\lambda)^{-1} \leq r_p.$$

Proof. Given that ϕ is symmetric with respect to 0, without loss of generality, we assume $\lambda > 0$. Then, we have

$$\frac{\lambda}{\phi(\lambda)} = \frac{1}{1 + 32\delta\lambda^{-1}} \cdot \frac{\lceil \log_2(\lambda/\delta) \rceil}{p_{\max}} \leq \frac{1}{p_{\max}} \cdot \frac{p+1}{1 + 2^{-(p-4)}} = r_p,$$

and

$$\frac{\lambda}{\phi(\lambda)} = \frac{1}{1 + 32\delta\lambda^{-1}} \cdot \frac{\lceil \log_2(\lambda/\delta) \rceil}{p_{\max}} > \frac{1}{p_{\max}} \cdot \frac{p+1}{1 + 2^{-(p-5)}} = l_p.$$

□

Lemma 4. $\lambda\phi(\lambda)^{-1}$ is monotonically increasing for $\lambda \in [-L_1, L_1]$.

Proof. Note that $\lambda \cdot \phi(\lambda)^{-1}$ is an odd function. Therefore, it suffices to prove that

$$\lambda \cdot \phi(\lambda)^{-1} = \frac{\lambda}{(32\delta + \lambda)} \cdot \frac{\lceil \log_2(\max\{\lambda, 2\delta\}/\delta) \rceil}{\lceil \log_2(L_1/\delta) \rceil}.$$

is increasing for $\lambda \in [0, L_1]$, where it suffices to check monotonicity of

$$\frac{\lambda \cdot \lceil \log_2(\max\{\lambda, 2\delta\}/\delta) \rceil}{32\delta + \lambda}.$$

First observe that the function is strictly increasing for any $\lambda \in [0, 2\delta)$. For any positive integer $p \geq 0$ and any $\lambda \in [2^p\delta, 2^{p+1}\delta)$, we have

$$\frac{\lambda \cdot \lceil \log_2(\max\{\lambda, 2\delta\}/\delta) \rceil}{32\delta + \lambda} = \frac{\lambda \cdot p}{32\delta + \lambda},$$

where

$$\frac{d}{d\lambda} \left(\frac{\lambda \cdot p}{32\delta + \lambda} \right) = p \cdot \frac{(32\delta + \lambda) - \lambda}{(32\delta + \lambda)^2} = \frac{p \cdot 32\delta}{(32\delta + \lambda)^2} > 0.$$

Thus, the function is strictly increasing on each interval $[2^p\delta, 2^{p+1}\delta)$. Moreover, at each boundary point $\lambda = 2^p\delta$, the function increases from $\frac{2^p\delta \cdot (p-1)}{32\delta + 2^p\delta}$ to $\frac{2^p\delta \cdot p}{32\delta + 2^p\delta}$. We can thus conclude that $\lambda\phi^{-1}(\lambda)$ is monotonically increasing. $\square$

Lemma 5. *The function ϕ defined in (3) satisfies*

$$\min_{\lambda \in (-\infty, +\infty)} \phi(\lambda) \geq 12\delta p_{\max}.$$

Proof. Since ϕ is symmetric with respect to 0, we only need to consider values of λ in $[0, +\infty)$. We analyze the function in pieces. For $\lambda \in [0, 2\delta)$, we have

$$\phi(\lambda) = (32\delta + \lambda) \cdot \frac{p_{\max}}{\lceil \log_2(2\delta/\delta) \rceil} = (32\delta + \lambda) \cdot p_{\max},$$

which is minimized at $\lambda = 0$, yielding

$$\phi(0) = 32\delta \cdot p_{\max}.$$

For $\lambda \in [2^k\delta, 2^{k+1}\delta)$ with $k \geq 1$, we have

$$\phi(\lambda) = (32\delta + \lambda) \cdot \frac{p_{\max}}{\lceil \log_2(\lambda/\delta) \rceil} = (32\delta + \lambda) \cdot \frac{p_{\max}}{k},$$

which is minimized at $\lambda = 2^k\delta$, giving

$$\phi(2^k\delta) = \frac{32 + 2^k}{k} \cdot \delta p_{\max}.$$

Hence, we have

$$\min_{\lambda \in (-\infty, +\infty)} \phi(\lambda) = \min \left\{ 32, \min_{k \geq 1} \frac{32 + 2^k}{k} \right\} \cdot \delta p_{\max} = 12\delta p_{\max}.$$

$\square$

Corollary 3. *For any symmetric $H \in \mathbb{R}^{d \times d}$, the matrix $\hat{H}$ defined in (2) satisfies $\hat{H} \succeq 12\delta p_{\max} \cdot I$ and $\hat{H}^{-1} \preceq (12\delta p_{\max})^{-1} I$.*

D Tools from Matrix Perturbation Theory and Extensions

In this section, we present several useful results from matrix perturbation theory that characterize how the eigenspaces of symmetric matrices change under perturbations. Recall the setting from Section 2.1, where we are given a δ -approximate Hessian $H_{x^{(0)}}$ at the point $x^{(0)}$, satisfying $\|H_{x^{(0)}} - \nabla^2 f(x^{(0)})\| \leq \delta$ and $-3\delta I \preceq H_{x^{(0)}} \preceq L_1 I$. Throughout this section, we denote $G := \nabla^2 f(x^{(0)})$. Our goal is to show that the spectra of $\hat{H}$ and $\hat{H}^{-1/2} G \hat{H}^{-1/2}$ can be partitioned into $\Theta(p_{\max})$ contiguous subsets such that, for any $p_{\min} \leq p \leq p_{\max}$, the principal angles between the eigenspaces spanned by the p -th spectral subsets of $\hat{H}$ and $\hat{H}^{-1/2} G \hat{H}^{-1/2}$ are small. We also derive several additional properties of these eigenspaces, which are used in the analysis in Appendix E.

D.1 Definition of Projectors

This part includes the definition of a series of projectors that we use to analyze the spectrum of $\hat{H}$ and $\hat{H}^{-1/2}G\hat{H}^{-1/2}$. Define

$$\hat{\Pi}_p^+ := \Pi_{(l_p, L_1]}(H_{x^{(0)}}) = \Pi_{(l_p, \infty)}(H_{x^{(0)}}), \quad \hat{\Pi}_p^- := \Pi_{[-2\delta, l_{p+1}]}(H_{x^{(0)}}), \quad (15)$$

and

$$\hat{\Pi}_p := \Pi_{(l_p, l_{p+1}]}(H_{x^{(0)}}). \quad (16)$$

Similarly, we define

$$\bar{\Pi}_p^+ := \Pi_{(\bar{l}_p, \infty)}(\hat{H}^{-1/2}G\hat{H}^{-1/2}), \quad \bar{\Pi}_p^- := \Pi_{(-\infty, \bar{l}_{p+1}]}(\hat{H}^{-1/2}G\hat{H}^{-1/2}), \quad (17)$$

and

$$\bar{\Pi}_p := \Pi_{(\bar{l}_p, \bar{l}_{p+1}]}(\hat{H}^{-1/2}G\hat{H}^{-1/2}). \quad (18)$$

Furthermore, we define

$$\bar{\Pi}_{\text{base}} := \bar{\Pi}_{p-1}^-. \quad (19)$$

D.2 The Davis-Kahan Theorem

In this subsection, we present the celebrated Davis–Kahan theorem, and provide an equivalent formulation that we use in our paper.

Definition 2 (Principal angles of subspaces). *Let $V, \tilde{V} \subset \mathbb{R}^n$ be k -dimensional subspaces. The principal angles $\theta_1(V, \tilde{V}), \dots, \theta_k(V, \tilde{V})$ between V and $\tilde{V}$ are defined recursively by*

$$(v_j, \tilde{v}_j) := \arg \max_{v_j \in V, \tilde{v}_j \in \tilde{V}} \langle v_j, \tilde{v}_j \rangle, \quad \theta_j(V, \tilde{V}) := \arccos \langle v_j, \tilde{v}_j \rangle,$$

subject to the constraint

$$\|v_j\| = \|\tilde{v}_j\| = 1, \quad \langle v_j, v_i \rangle = 0, \quad \langle \tilde{v}_j, \tilde{v}_i \rangle = 0, \quad \forall 1 \leq i < j.$$

Lemma 6 (Davis-Kahan Theorem, see e.g., Theorem 1 of [38]). *Let $M, \tilde{M} \in \mathbb{R}^{d \times d}$ be two symmetric matrices satisfying $\|M - \tilde{M}\| \leq \xi$ for some $\xi > 0$. For any $a < b$, we use $\mathcal{S} = \{v_1, \dots, v_k\}$ and $\tilde{\mathcal{S}} = \{\tilde{v}_1, \dots, \tilde{v}_{\tilde{k}}\}$ to denote the set of normalized eigenvectors of M and $\tilde{M}$ associated with eigenvalues contained in the interval $[a, b]$ and $[a - \xi, b + \xi]$, respectively, and denote*

$$V := \text{span}(\mathcal{S}), \quad \tilde{V} := \text{span}(\tilde{\mathcal{S}}).$$

Then, if the remaining eigenvalues of M lie outside the interval $[a - \gamma, b + \gamma]$, we have $k = \tilde{k}$ and

$$\|\sin \Theta(V, \tilde{V})\| \leq \frac{\xi}{\gamma},$$

where

$$\sin \Theta(V, \tilde{V}) := \text{diag}(\sin \theta_1(V, \tilde{V}), \dots, \sin \theta_k(V, \tilde{V}))^\top \quad (20)$$

Lemma 7 (Theorem I.5.5 of [35]). *Let $V, \tilde{V} \subset \mathbb{R}^n$ be k -dimensional subspaces and let $\Pi, \tilde{\Pi}$ be their projectors. Then, we have*

$$\|\Pi - \tilde{\Pi}\| = \|\sin \Theta(V, \tilde{V})\| = \sin \theta_1(V, \tilde{V}),$$

where $\sin \Theta(V, \tilde{V})$ is defined in (20).

Lemma 8 (Equivalent form of the Davis-Kahan Theorem). *Let $M, \tilde{M} \in \mathbb{R}^{d \times d}$ be two symmetric matrices satisfying $\|M - \tilde{M}\| \leq \xi$ for some $\xi > 0$. For any $a < b$ and $\gamma > \xi$, if there are no eigenvalues of M in intervals $[a - \gamma, a]$ and $[b, b + \gamma]$, we have $k = \tilde{k}$ and*

$$\|\Pi_{[a, b]}(M) - \Pi_{[a - \gamma, b + \gamma]}(\tilde{M})\| = \left\| \sum_{j=1}^k v_j v_j^\top - \sum_{j=1}^k \tilde{v}_j \tilde{v}_j^\top \right\| \leq \frac{\xi}{\gamma}.$$

Proof. The proof follows by combining Lemma 6 and Lemma 7. □

Intuitively, Lemma 6 and Lemma 8 states that a small perturbation to a symmetric matrix leads to only a small change in its eigenspaces, provided the corresponding eigenvalues are well separated from the other eigenvalues.

D.3 Properties of $\hat{\Pi}_p^+$ and $\bar{\Pi}_p^+$

Proposition 1. For any positive integer $p \geq 5$, we have

$$\|\hat{\Pi}_p^+ - \bar{\Pi}_p^+\| \leq 2^{2-p/2} \sqrt{p}.$$

Before proving Proposition 1, we first present the following two lemmas.

Lemma 9. For any positive integer $p \geq 5$, denote

$$\tilde{G} := \hat{\Pi}_p^+ H_{x(0)} \hat{\Pi}_p^+ + (I - \hat{\Pi}_p^+) G (I - \hat{\Pi}_p^+)$$

with $\hat{\Pi}_p^+$ defined in (15). Then, we have

$$\|\hat{H}^{-1/2} (G - \tilde{G}) \hat{H}^{-1/2}\| \leq \xi_p,$$

where $\hat{H}$ is defined in (2).

Proof. Given that $H_{x(0)}$ and $\hat{H}$ have the same set of eigenvectors, we have

$$\hat{\Pi}_p^+ \hat{H} (I - \hat{\Pi}_p^+) = (I - \hat{\Pi}_p^+) \hat{H} \hat{\Pi}_p^+ = 0.$$

Then, $\tilde{G}$ can also be written as

$$\tilde{G} = H_{x(0)} - (I - \hat{\Pi}_p^+) H_{x(0)} (I - \hat{\Pi}_p^+) + (I - \hat{\Pi}_p^+) G (I - \hat{\Pi}_p^+) = H_{x(0)} + (I - \hat{\Pi}_p^+) (G - H_{x(0)}) (I - \hat{\Pi}_p^+),$$

which implies

$$\begin{aligned} G - \tilde{G} &= (G - H_{x(0)}) - (I - \hat{\Pi}_p^+) (G - H_{x(0)}) (I - \hat{\Pi}_p^+) \\ &= \hat{\Pi}_p^+ (G - H_{x(0)}) + (G - H_{x(0)}) \hat{\Pi}_p^+ - \hat{\Pi}_p^+ (G - H_{x(0)}) \hat{\Pi}_p^+. \end{aligned}$$

Therefore,

$$\begin{aligned} \|\hat{H}^{-1/2} (G - \tilde{G}) \hat{H}^{-1/2}\| &\leq \|\hat{H}^{-1/2} \hat{\Pi}_p^+ (G - H_{x(0)}) \hat{H}^{-1/2}\| + \|\hat{H}^{-1/2} (G - H_{x(0)}) \hat{\Pi}_p^+ \hat{H}^{-1/2}\| \\ &\quad + \|\hat{H}^{-1/2} \hat{\Pi}_p^+ (G - H_{x(0)}) \hat{\Pi}_p^+ \hat{H}^{-1/2}\|. \end{aligned}$$

From the definitions of $\hat{H}$ and ϕ in equations (2) and (3), it follows that $\phi(\lambda) \geq 12\delta$ for all $\lambda \in [-L, L]$, yielding

$$\|\hat{H}^{-1/2}\| \leq \sqrt{\frac{1}{12\delta p_{\max}}} < \frac{1}{3} \sqrt{\frac{1}{\delta p_{\max}}}$$

and

$$\|\hat{\Pi}_p^+ \hat{H}^{-1/2}\| = \|\hat{H}^{-1/2} \hat{\Pi}_p^+\| \leq \sqrt{\frac{p}{2^p \delta p_{\max}}}.$$

Combining these bounds gives

$$\|\hat{H}^{-1/2} \hat{\Pi}_p^+ (G - H_{x(0)}) \hat{H}^{-1/2}\| = \|\hat{H}^{-1/2} (G - H_{x(0)}) \hat{\Pi}_p^+ \hat{H}^{-1/2}\| \leq \frac{\sqrt{p}}{3 \times 2^{p/2} p_{\max}},$$

and

$$\|\hat{H}^{-1/2} \hat{\Pi}_p^+ (G - H_{x(0)}) \hat{\Pi}_p^+ \hat{H}^{-1/2}\| \leq \frac{\sqrt{p}}{2^p p_{\max}}.$$

Thus, we can conclude that

$$\|\hat{H}^{-1/2} (G - \tilde{G}) \hat{H}^{-1/2}\| \leq \frac{p}{2^{p/2} p_{\max}} = \xi_p.$$

□

Lemma 10. For any positive integer $p \geq 5$, we have

$$\hat{\Pi}_p^+ \hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2} \hat{\Pi}_p^+ \succeq l_p \cdot \hat{\Pi}_p^+$$

and

$$\|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} G \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| \leq \frac{1}{12 p_{\max}} + r_{p-1},$$

where $\hat{\Pi}_p^+$ and $\hat{H}$ are defined in (15) and (2), respectively.

Proof. Given that $H_{x(0)}$ and $\hat{H}$ share the same eigenvectors, the minimal eigenvalue of $\hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2}$ restricted to the subspace projected by $\hat{\Pi}_p^+$ is bounded below by

$$\min_{\lambda > 2^p \delta} \lambda \phi^{-1}(\lambda) > l_p,$$

as established in Lemma 3. Consequently, we have

$$\hat{\Pi}_p^+ \hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2} \hat{\Pi}_p^+ \succeq l_p \cdot \hat{\Pi}_p^+.$$

By the triangle inequality, it follows that

$$\begin{aligned} \|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} G \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| &\leq \|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| \\ &\quad + \|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} (G - H_{x(0)}) \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\|. \end{aligned}$$

By the definition of $\hat{H}$ in (2) and Lemma 3, we have the bound

$$\|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| \leq \frac{1}{p_{\max}} \cdot \frac{p}{1 + 2^{-(p-5)}} = r_{p-1}.$$

Moreover, since

$$\|\hat{H}^{-1/2}\| \leq \sqrt{\frac{1}{12\delta p_{\max}}},$$

we obtain

$$\begin{aligned} \|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} (G - H_{x(0)}) \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| &\leq \|\hat{H}^{-1/2} (G - H_{x(0)}) \hat{H}^{-1/2}\| \\ &\leq \|\hat{H}^{-1/2}\|^2 \cdot \|G - H_{x(0)}\| \leq \frac{1}{12p_{\max}}. \end{aligned}$$

By combining these bounds we can conclude that

$$\|(I - \hat{\Pi}_p^+) \hat{H}^{-1/2} G \hat{H}^{-1/2} (I - \hat{\Pi}_p^+)\| \leq \frac{1}{12p_{\max}} + r_{p-1}.$$

□

Proof of Proposition 1. Define

$$\tilde{G} := \hat{\Pi}_p^+ H_{x(0)} \hat{\Pi}_p^+ + (I - \hat{\Pi}_p^+) G (I - \hat{\Pi}_p^+).$$

Since $\hat{H}$ and $\hat{\Pi}_p^+$ share the same set of eigenvectors, in the basis $\{\hat{h}_1, \dots, \hat{h}_d\}$, where the eigenvectors are arranged in descending order according to their eigenvalues, the matrix $\hat{H}^{-1/2} \tilde{G} \hat{H}^{-1/2}$ takes the following block-diagonal form:

$$\hat{H}^{-1/2} \tilde{G} \hat{H}^{-1/2} = \begin{bmatrix} \hat{\Pi}_p^+ \hat{H}^{-1/2} H_{x(0)} \hat{H}^{-1/2} \hat{\Pi}_p^+ & 0 \\ 0 & (I - \hat{\Pi}_p^+) \hat{H}^{-1/2} G \hat{H}^{-1/2} (I - \hat{\Pi}_p^+) \end{bmatrix}.$$

By Lemma 10, the top left block has eigenvalues bounded below by l_p , while the bottom right block has eigenvalues bounded above by $r_{p-1} + \frac{1}{12\lceil \log_2(L/\delta) \rceil}$, with eigenvalue gap

$$l_p - r_{p-1} - \frac{1}{12p_{\max}} \geq \frac{1}{3p_{\max}} > \xi_{p-1}.$$

Additionally, by Lemma 9, we have the bound

$$\|\hat{H}^{-1/2} G \hat{H}^{-1/2} - \hat{H}^{-1/2} \tilde{G} \hat{H}^{-1/2}\| \leq \xi_p.$$

We can then conclude by applying Lemma 8, which gives

$$\|\hat{\Pi}_p^+ - \bar{\Pi}_p^+\| \leq 2^{2-p/2} \sqrt{p}.$$

□

D.4 Properties of $\hat{\Pi}_p^-, \hat{\Pi}_p$ and $\bar{\Pi}_p^-, \bar{\Pi}_p$

Lemma 11. For any positive integer $p \geq 5$, we have

$$\|\hat{\Pi}_p^- - \bar{\Pi}_p^-\| \leq 2^{2-p/2} \sqrt{p}.$$

Proof. Given that

$$\hat{\Pi}_p^- + \hat{\Pi}_{p+1}^+ = \bar{\Pi}_p^- + \bar{\Pi}_{p+1}^+ = I,$$

we have

$$\|\hat{\Pi}_p^- - \bar{\Pi}_p^-\| = \|\hat{\Pi}_p^+ - \bar{\Pi}_{p+1}^+\| \leq 2^{2-p/2}.$$

□

Lemma 12. For any positive integer $p > 1$, we have

$$\|\hat{\Pi}_p - \bar{\Pi}_p\| \leq 2^{3-p/2} \sqrt{p}.$$

Proof. Given that

$$\tilde{\Pi}_{p-1}^- + \hat{\Pi}_p + \hat{\Pi}_{p+1}^+ = I, \quad \bar{\Pi}_p^- + \bar{\Pi}_p + \bar{\Pi}_{p+1}^+ = I,$$

we have

$$\|\hat{\Pi}_p - \bar{\Pi}_p\| \leq \|\tilde{\Pi}_{p-1}^- - \bar{\Pi}_{p-1}^-\| + \|\hat{\Pi}_{p+1}^+ - \bar{\Pi}_{p+1}^+\| \leq 2^{2-p/2} \sqrt{p} + 2^{2-p/2} \sqrt{p} \leq 2^{3-p/2} \sqrt{p}.$$

□

Lemma 13. For any positive integer $p \geq 5$ and any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p \hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p \hat{v}\|$$

and

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p^- \hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p^- \hat{v}\|.$$

Proof. For the first inequality, observe that

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p \hat{v}\| \leq \sum_{q=p+1}^{p_{\max}} \|\hat{H}^{1/2} \hat{\Pi}_q \bar{\Pi}_p \hat{v}\|,$$

where for each $q > p$ we have

$$\|\hat{H}^{1/2} \hat{\Pi}_q \bar{\Pi}_p \hat{v}\| \leq 2^{2-q/2} \sqrt{q} \|\hat{H}^{1/2} \hat{\Pi}_q\| \cdot \|\bar{\Pi}_p \hat{v}\| \leq 2^{2-q/2} \sqrt{q\phi(2^{q+1}\delta)} \cdot \|\bar{\Pi}_p \hat{v}\|$$

by Lemma 11. Summing over q gives

$$\begin{aligned} \sum_{q=p+1}^{p_{\max}} 2^{2-p/2} \cdot \phi(2^{q+1}\delta) &= \sum_{q=p+1}^{p_{\max}} 2^{2-q/2} \cdot \sqrt{(32+2^q) \cdot \delta p_{\max}} \\ &\leq \sqrt{\delta} \cdot p_{\max}. \end{aligned}$$

Therefore,

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p \hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p \hat{v}\|.$$

The proof of the second inequality is similar. Note that

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p^- \hat{v}\| \leq \sum_{q=p+1}^{p_{\max}} \|\hat{H}^{1/2} \hat{\Pi}_q \bar{\Pi}_p^- \hat{v}\|,$$

where for each $q > p$ we have

$$\|\hat{H}^{1/2} \hat{\Pi}_q \bar{\Pi}_p^- \hat{v}\| \leq 2^{2-q/2} \sqrt{q} \|\hat{H}^{1/2} \hat{\Pi}_q\| \cdot \|\bar{\Pi}_p^- \hat{v}\| \leq 2^{2-p/2} \sqrt{q\phi(2^{q+1}\delta)} \cdot \|\bar{\Pi}_p^- \hat{v}\|$$

by Lemma 11. Summing over q gives

$$\sum_{q=p+1}^{p_{\max}} 2^{2-p/2} \cdot \phi(2^{q+1}\delta) = \sum_{q=p+1}^{p_{\max}} 2^{2-q/2} \cdot \sqrt{(32+2^q) \cdot \delta p_{\max}} \leq \sqrt{\delta} \cdot p_{\max}.$$

Therefore,

$$\|\hat{H}^{1/2} \hat{\Pi}_{p+1}^+ \bar{\Pi}_p^- \hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p^- \hat{v}\|.$$

□

Proposition 2. For any positive integer $p \geq 5$ and any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq 2^{p/2}\sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p\hat{v}\|$$

and

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \geq \|\bar{\Pi}_p\hat{v}\|$$

Moreover, we have

$$\|\hat{H}^{-1/2}\bar{\Pi}_p\hat{v}\| \geq \frac{1}{2\sqrt{\phi(2^{p+1}\delta)}} \|\bar{\Pi}_p\hat{v}\|.$$

Proof. For the first inequality, note that

$$\bar{\Pi}_p\hat{v} = \hat{\Pi}_p^-\bar{\Pi}_p\hat{v} + \sum_{q=p+1}^{p_{\max}} \hat{\Pi}_q\bar{\Pi}_p\hat{v},$$

which gives

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p\hat{v}\| + \sum_{q=p+1}^{p_{\max}} \|\hat{H}^{1/2}\hat{\Pi}_q\bar{\Pi}_p\hat{v}\|,$$

where

$$\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p\hat{v}\| \leq \|\hat{H}^{1/2}\hat{\Pi}_p^-\| \cdot \|\bar{\Pi}_p\hat{v}\| \leq \sqrt{\phi(2^{p+1}\delta)} \cdot \|\bar{\Pi}_p\hat{v}\|$$

and

$$\sum_{q=p+1}^{p_{\max}} \|\hat{H}^{1/2}\hat{\Pi}_q\bar{\Pi}_p\hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p\hat{v}\|$$

by Lemma 13. Hence,

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq 2^{p/2}\sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p\hat{v}\|.$$

The second inequality follows from

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \geq \|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| \geq \|\hat{H}^{1/2}\hat{\Pi}_p\| \cdot \|\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| \geq \frac{\sqrt{\phi(2^p\delta)}}{2} \|\bar{\Pi}_p\hat{v}\|$$

As for the third inequality,

$$\bar{\Pi}_p\hat{v} = \hat{\Pi}_p\bar{\Pi}_p\hat{v} + \hat{\Pi}_{p-1}^-\bar{\Pi}_p\hat{v} + \hat{\Pi}_{p+1}^+\bar{\Pi}_p\hat{v},$$

which leads to

$$\|\hat{H}^{-1/2}\bar{\Pi}_p\hat{v}\| \geq \|\hat{H}^{-1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| \geq \frac{1}{\sqrt{\phi(2^{p+1}\delta)}} \cdot \|\hat{\Pi}_p\bar{\Pi}_p\hat{v}\|,$$

where

$$\|\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| \geq \|\bar{\Pi}_p\bar{\Pi}_p\hat{v}\| - \|(\hat{\Pi}_p - \bar{\Pi}_p)\bar{\Pi}_p\hat{v}\| \geq \frac{1}{2} \|\bar{\Pi}_p\hat{v}\|,$$

by which we can conclude that

$$\|\hat{H}^{-1/2}\bar{\Pi}_p\hat{v}\| \geq \frac{1}{2\sqrt{\phi(2^{p+1}\delta)}} \|\bar{\Pi}_p\hat{v}\|.$$

□

D.5 The Connection between $\widehat{H}^{1/2}\hat{v}$ and $\widehat{H}^{1/2}\overline{\Pi}_p\hat{v}$

In this subsection, we prove the following result.

Proposition 3. *For any positive integer $p \geq p_{\min}$ and any $\hat{v} \in \mathbb{R}^d$, we have*

$$\|\widehat{H}^{1/2}\hat{v}\| \geq \frac{2^{-5}}{1 + 2^{-p/4}\sqrt{p_{\max}}} \cdot \|\widehat{H}^{1/2}\overline{\Pi}_p\hat{v}\|. \quad (21)$$

Denote $v_p := \widehat{H}^{1/2}\overline{\Pi}_p\hat{v}$, $v_p^- := \widehat{H}^{1/2}\overline{\Pi}_{p-1}^-\hat{v}$, and $v_p^+ := \widehat{H}^{1/2}\overline{\Pi}_{p+1}^+\hat{v}$. Then, we have $\widehat{H}^{1/2}\hat{v} = v_p + v_p^- + v_p^+$. If $v_p = 0$, (21) holds directly. Hence, we only need to prove the case where $v_p \neq 0$.

The following inequalities are useful for proving Proposition 3.

Lemma 14. *For any integer $p \geq 5$ and any $\hat{v} \in \mathbb{R}^d$, the following inequalities hold.*

1. $\|\widehat{H}^{1/2}\widehat{\Pi}_p\overline{\Pi}_p\hat{v}\| \geq \sqrt{\phi(2^p\delta)}\|\overline{\Pi}_p\hat{v}\|/2$;
2. $\|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\overline{\Pi}_p\hat{v}\| \leq 2^{2-p/2}\sqrt{p\phi(2^p\delta)}\|\overline{\Pi}_p\hat{v}\|$;
3. $\|\widehat{H}^{1/2}\widehat{\Pi}_p\overline{\Pi}_{p-1}^-\hat{v}\| \leq 2^{2-p/2}\sqrt{p\phi(2^{p+1}\delta)}\|\overline{\Pi}_{p-1}^-\hat{v}\|$;
4. $\|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\overline{\Pi}_{p-1}^-\hat{v}\| \geq \sqrt{\delta \cdot p_{\max}} \cdot \|\overline{\Pi}_{p-1}^-\hat{v}\|$;
5. $\|\widehat{H}^{1/2}\widehat{\Pi}_{p+1}^+\overline{\Pi}_{p-1}^-\hat{v}\| \leq \sqrt{\delta \cdot p_{\max}} \cdot \|\overline{\Pi}_{p-1}^-\hat{v}\|$;
6. $\|\widehat{H}^{1/2}\widehat{\Pi}_p\overline{\Pi}_{p+1}^+\hat{v}\| \leq 2^{2-p/2}\sqrt{p\phi(2^{p+1}\delta)}\|\overline{\Pi}_{p+1}^+\hat{v}\|$;
7. $\|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\overline{\Pi}_{p+1}^+\hat{v}\| \leq 2^{1-p/2}\sqrt{p\phi(2^p\delta)}\|\overline{\Pi}_{p+1}^+\hat{v}\|$;
8. $\|\widehat{H}^{1/2}\widehat{\Pi}_{p+1}^+\overline{\Pi}_{p+1}^+\hat{v}\| \geq \sqrt{\phi(2^{p+1}\delta)}\|\overline{\Pi}_{p+1}^+\hat{v}\|/2$.

Proof. Proof of the first entry:

$$\begin{aligned} \|\widehat{H}^{1/2}\widehat{\Pi}_p\overline{\Pi}_p\hat{v}\| &\geq \sqrt{\phi(2^p\delta)}\|\widehat{\Pi}_p\overline{\Pi}_p\hat{v}\| \\ &\geq \sqrt{\phi(2^p\delta)}(\|\overline{\Pi}_p\overline{\Pi}_p\hat{v}\| - \|(\widehat{\Pi}_p - \overline{\Pi}_p)\overline{\Pi}_p\hat{v}\|) \geq \frac{\sqrt{\phi(2^p\delta)}}{2}\|\overline{\Pi}_p\hat{v}\| \end{aligned}$$

by Lemma 12.

Proof of the second entry:

$$\begin{aligned} \|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\overline{\Pi}_p\hat{v}\| &\leq \|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\| \cdot \|\widehat{\Pi}_{p-1}^-\overline{\Pi}_p\hat{v}\| \\ &\leq \sqrt{\phi(2^p\delta)}\|\widehat{\Pi}_{p-1}^-\overline{\Pi}_p\hat{v}\| \leq 2^{2-p/2}\sqrt{p \cdot \phi(2^p\delta)}\|\overline{\Pi}_p\hat{v}\| \end{aligned}$$

by Lemma 11.

Proof of the third entry:

$$\begin{aligned} \|\widehat{H}^{1/2}\widehat{\Pi}_p\overline{\Pi}_{p-1}^-\hat{v}\| &\leq \|\widehat{H}^{1/2}\widehat{\Pi}_p\| \cdot \|\widehat{\Pi}_p\overline{\Pi}_{p-1}^-\hat{v}\| \\ &\leq \sqrt{\phi(2^{p+1}\delta)} \cdot \|\widehat{\Pi}_p\overline{\Pi}_{p-1}^-\hat{v}\| \\ &\leq 2^{2-p/2}\sqrt{p\phi(2^{p+1}\delta)}\|\overline{\Pi}_{p-1}^-\hat{v}\|, \end{aligned}$$

by Lemma 12.

Proof of the fourth entry:

$$\begin{aligned} \|\widehat{H}^{1/2}\widehat{\Pi}_{p-1}^-\overline{\Pi}_{p-1}^-\hat{v}\| &\geq \frac{1}{\|\widehat{H}^{-1/2}\|}\|\widehat{\Pi}_{p-1}^-\overline{\Pi}_{p-1}^-\hat{v}\| \\ &\geq \sqrt{12\delta\lceil\log_2(L/\delta)\rceil}(\|\overline{\Pi}_{p-1}^-\hat{v}\| - \|(\widehat{\Pi}_{p-1}^- - \overline{\Pi}_{p-1}^-)\overline{\Pi}_{p-1}^-\hat{v}\|) \\ &\geq \sqrt{\delta\lceil\log_2(L/\delta)\rceil}\|\overline{\Pi}_{p-1}^-\hat{v}\| \end{aligned}$$

by Lemma 11.

Proof of the fifth entry:

$$\|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_{p-1}^-\hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_{p-1}^-\hat{v}\|$$

by Lemma 13.

Proof of the sixth entry:

$$\begin{aligned} \|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_{p+1}^+\hat{v}\| &\leq \|\hat{H}^{1/2}\hat{\Pi}_p\| \cdot \|(\hat{\Pi}_p - \bar{\Pi}_p)\bar{\Pi}_{p+1}^+\hat{v}\| \\ &\leq 2^{2-p/2} \sqrt{p\phi(2^{p+1}\delta)} \|\bar{\Pi}_{p+1}^+\hat{v}\| \end{aligned}$$

by Lemma 12.

Proof of the seventh entry:

$$\begin{aligned} \|\hat{H}^{1/2}\hat{\Pi}_{p-1}^-\bar{\Pi}_{p+1}^+\hat{v}\| &\leq \|\hat{H}^{1/2}\hat{\Pi}_{p-1}^-\| \cdot \|(\hat{\Pi}_{p-1}^- - \bar{\Pi}_{p-1}^-)\bar{\Pi}_{p+1}^+\hat{v}\| \\ &\leq 2^{1-p/2} \sqrt{p\phi(2^p\delta)} \|\bar{\Pi}_{p+1}^+\hat{v}\| \end{aligned}$$

by Lemma 11.

Proof of the eighth entry:

$$\begin{aligned} \|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_{p+1}^+\hat{v}\| &\geq \sqrt{\phi(2^{p+1}\delta)} (\|\bar{\Pi}_{p+1}^+\hat{v}\| - \|(\hat{\Pi}_{p+1}^+ - \bar{\Pi}_{p+1}^+)\bar{\Pi}_{p+1}^+\hat{v}\|) \\ &\geq \frac{\sqrt{\phi(2^{p+1}\delta)}}{2} \|\bar{\Pi}_{p+1}^+\hat{v}\| \end{aligned}$$

by Proposition 1. □

Lemma 15. For any positive integer $p \geq p_{\min}$ and any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\| \geq \sin\left(\frac{\pi}{p_{\min}}\right) \cdot \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\|.$$

Proof. Note that

$$\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v} = \hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v} + \hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_{p-1}^-\hat{v}$$

If either $\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}$ or $\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_{p-1}^-\hat{v}$ equals 0, the inequality holds directly. Otherwise, by entries 1 through 4 of Lemma 14, we have

$$\arccos\left(\frac{\langle \hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}, \hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_{p-1}^-\hat{v} \rangle}{\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\| \cdot \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_{p-1}^-\hat{v}\|}\right) \geq \frac{\pi}{16},$$

which leads to

$$\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\| \geq \sin(\pi/16) \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\|.$$

□

Lemma 16. For any positive integer $p \geq 5$ and any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_p^-\hat{v}\| \leq \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\|.$$

Proof. By Lemma 13 we have

$$\|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_p^-\hat{v}\| \leq \sqrt{\delta} \cdot p_{\max} \|\bar{\Pi}_p^-\hat{v}\|.$$

Moreover, by entry 4 of Lemma 14 we have

$$\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\| \geq \sqrt{\delta \cdot p_{\max}} \cdot \|\bar{\Pi}_p^-\hat{v}\|,$$

which leads to

$$\|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_p^-\hat{v}\| \leq \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\|.$$

□

Lemma 17. For any positive integer $p \geq 5$ and $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq (2 + 2^{-p/4}\sqrt{p_{\max}})\|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\|.$$

Proof.

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq \|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| + \|\hat{H}^{1/2}\hat{\Pi}_{p-1}^-\bar{\Pi}_p\hat{v}\| + \|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_p\hat{v}\|,$$

where

$$\|\hat{H}^{1/2}\hat{\Pi}_{p-1}^-\bar{\Pi}_p\hat{v}\| \leq 2^{2-p/2}\sqrt{p\phi(2^p\delta)}\|\bar{\Pi}_p\hat{v}\| \leq 2^{2-(p-\log p)/2}\|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\|$$

by the first two inequalities in Lemma 14, and

$$\begin{aligned}\|\hat{H}^{1/2}\hat{\Pi}_{p+1}^+\bar{\Pi}_p\hat{v}\| &\leq \sqrt{\delta} \cdot p_{\max} \cdot \|\bar{\Pi}_p\hat{v}\| \\ &\leq 2\sqrt{\frac{\delta}{\phi(2^p\delta)}} \cdot p_{\max} \cdot \|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\| \\ &\leq 2^{-p/4}\sqrt{p_{\max}}\|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\|\end{aligned}$$

by Proposition 2 and the first inequality of Lemma 14. We can therefore conclude that

$$\|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\| \leq (2 + 2^{-p/4}\sqrt{p_{\max}})\|\hat{H}^{1/2}\hat{\Pi}_p\bar{\Pi}_p\hat{v}\|.$$

□

Equipped with these results, we are ready to prove Proposition 3.

Proof of Proposition 3. Note that $\hat{H}^{1/2}\hat{v} = \hat{H}^{1/2}\bar{\Pi}_p^-\hat{v} + \hat{H}^{1/2}\bar{\Pi}_{p+1}^+\hat{v}$. By entries 6 through 8 of Lemma 14, we have

$$\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_{p+1}^+\hat{v}\| \leq 2^{4-p/2}\|\hat{H}^{1/2}\bar{\Pi}_{p+1}^+\hat{v}\|,$$

indicating that the angle between $\hat{H}^{1/2}\bar{\Pi}_{p+1}^+\hat{v}$ and the subspace projected by $\hat{\Pi}_p^-$ is at most $\arcsin(2^{4-p/2})$. On the other hand, by Lemma 16 we know that the angle between $\hat{H}^{1/2}\bar{\Pi}_p^-\hat{v}$ and the subspace projected by $\hat{\Pi}_p^-$ is at least $\pi/4$, which leads to

$$\begin{aligned}\|\hat{H}^{1/2}\hat{v}\| &= \|\hat{H}^{1/2}\bar{\Pi}_p^-\hat{v} + \hat{H}^{1/2}\bar{\Pi}_{p+1}^+\hat{v}\| \geq \frac{1}{2}\|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p^-\hat{v}\| \\ &\geq \frac{1}{2}\sin\left(\frac{\pi}{p_{\min}}\right) \cdot \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p\hat{v}\| \geq 2^{-4} \cdot \|\hat{H}^{1/2}\hat{\Pi}_p^-\bar{\Pi}_p\hat{v}\|\end{aligned}$$

by Lemma 15. Then by Lemma 17, we can conclude that

$$\|\hat{H}^{1/2}\hat{v}\| \geq \frac{2^{-5}}{1 + 2^{-p/4}\sqrt{p_{\max}}} \cdot \|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\|.$$

□

D.6 Properties of $\bar{\Pi}_{\text{base}}$

Lemma 18. For any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\bar{\Pi}_{\text{base}}\hat{v}\| \leq 40p_{\max}^{3/2}\|\hat{H}^{1/2}\hat{v}\|.$$

Proof. Given that

$$\hat{H}^{1/2}\hat{v} = \hat{H}^{1/2}\bar{\Pi}_{\text{base}}\hat{v} + \sum_{p=p_{\min}}^{p_{\max}} \hat{H}^{1/2}\bar{\Pi}_p\hat{v},$$

we have

$$\|\hat{H}^{1/2}\bar{\Pi}_{\text{base}}\hat{v}\| \leq \|\hat{H}^{1/2}\hat{v}\| + \sum_{p=p_{\min}}^{p_{\max}} \|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\|,$$

where

$$\|\hat{H}^{1/2}\hat{v}\| \geq \frac{2^{-5}}{1 + 2^{-p/4}\sqrt{p_{\max}}} \cdot \|\hat{H}^{1/2}\bar{\Pi}_p\hat{v}\|, \quad \forall p_{\min} \leq p \leq p_{\max}$$

by Proposition 3. We can thus conclude that

$$\|\hat{H}^{1/2}\bar{\Pi}_{\text{base}}\hat{v}\| \leq (1 + p_{\max}(32 + 2\sqrt{p_{\max}}))\|\hat{H}^{1/2}\hat{v}\| \leq 40p_{\max}^{3/2}\|\hat{H}^{1/2}\hat{v}\|.$$

□

Corollary 4. For any $\hat{v} \in \mathbb{R}^d$, we have

$$\|\hat{H}^{1/2}\bar{\Pi}_{\text{base}}\hat{v}\| \leq p_{\max}\sqrt{\delta} \cdot \|\bar{\Pi}_{\text{base}}\hat{v}\|.$$

Proof. The proof follows from Lemma 13 by noticing that

$$\bar{\Pi}_{\text{base}} = \bar{\Pi}_{p_{\min}-1}^-, \quad \bar{\Pi}_{\text{base}} + \bar{\Pi}_{p_{\min}}^+ = I$$

given the definition of $\mathcal{S}_{\text{base}}$ in (30). □

E Analysis of Algorithm 1

E.1 Quadratic Approximation of f Near $x^{(0)}$

Given that f is L_2 -Hessian Lipschitz, in the neighborhood of $x^{(0)}$, it is close to $T_{x^{(0)}}^2(x)$, is the 2nd order Taylor approximation of f evaluated at $x^{(0)}$. Throughout this section, we denote

$$g(x) := T_{x^{(0)}}^2(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle + \frac{1}{2}(x - x^{(0)})^\top \nabla^2 f(x^{(0)})(x - x^{(0)}). \quad (22)$$

Similarly to the definition of $\hat{f}$, we define $\hat{g} := (\hat{H}^{-1/2}\hat{x})$, which satisfies

$$\nabla \hat{g}(\hat{x}) = \hat{H}^{-1/2} \cdot \nabla g(\hat{H}^{-1/2}\hat{x}),$$

and

$$\bar{H} := \nabla^2 \hat{g}(\hat{x}) = \hat{H}^{-1/2} \cdot \nabla^2 g(\hat{H}^{-1/2}\hat{x}) \cdot \hat{H}^{-1/2} = \hat{H}^{-1/2} \cdot \nabla^2 f(x^{(0)}) \cdot \hat{H}^{-1/2},$$

For any iteration k , we define

$$\iota^{(k)} := \nabla f(y^{(k)}) - \nabla g(y^{(k)}), \quad \hat{\iota}^{(k)} := \nabla \hat{f}(\hat{y}^{(k)}) - \nabla \hat{g}(\hat{y}^{(k)}).$$

Then, we have $\iota^{(k)} = \hat{H}^{1/2}\hat{\iota}^{(k)}$.

E.2 Movement Bounds of the Iterates

In the case where Line 7 is triggered during Algorithm 1, we denote

$$\mathcal{K} := \operatorname{argmin}_k \left\{ k \sum_{t=0}^{k-1} \|x^{(t+1)} - x^{(t)}\|^2 > B^2 \right\}.$$

Otherwise, we denote $\mathcal{K} = K + 1$.

Lemma 19. For any iteration $k < \mathcal{K}$, we have

1. $k \sum_{t=1}^{k-1} \|x^{(t+1)} - x^{(t)}\|^2 < B^2$;
2. $\|x^{(k)} - x^{(0)}\| \leq B$;

3. $\|y^{(k)} - x^{(0)}\| \leq 2B$;
4. $\|\iota^{(k)}\| = \|\hat{H}^{1/2}\hat{\iota}^{(k)}\| \leq 2L_2B^2$

Proof. The first entry follows from Corollary 3:

$$\begin{aligned} k \sum_{k=1}^{k-1} \|x^{(k+1)} - x^{(k)}\|^2 &< k \sum_{t=1}^{k-1} (\|\hat{H}^{-1/2}\| \cdot \|\hat{H}^{1/2}(x^{(t+1)} - x^{(t)})\|)^2 \\ &< \frac{k}{12\delta p_{\max}} \sum_{t=1}^{k-1} \|\hat{H}^{1/2}(x^{(t+1)} - x^{(t)})\|^2 < B^2. \end{aligned}$$

Then by Cauchy-Schwartz, we have

$$\|x^{(k)} - x^{(0)}\| \leq \sqrt{k \sum_{t=1}^{k-1} \|\hat{H}^{1/2}(x^{(t+1)} - x^{(t)})\|^2} \leq B,$$

which leads to

$$\|y^{(k)} - x^{(0)}\| \leq \|x^{(k)} - x^{(0)}\| + (1 - \theta)\|x^{(k)} - x^{(k-1)}\| \leq 2B.$$

Since f is L_2 -Hessian Lipschitz, we can further derive that

$$\|\iota^{(k)}\| = \|\hat{H}^{1/2}\hat{\iota}^{(k)}\| \leq \frac{1}{2}L_2\|y^{(k)} - x^{(0)}\|^2 \leq 2L_2B^2.$$

□

Lemma 20. Let $\eta \leq 1/4$. In the case where the “if condition” in Line 7 of Algorithm 1 is triggered, we have $\|x^{(\mathcal{K})} - x^{(0)}\| \leq 7B$.

Proof. By Lemma 19, for any $k < \mathcal{K}$ we have $\|x^{(k)} - x^{(0)}\| \leq B$ and $\|y^{(k)} - x^{(0)}\| \leq 2B$. Hence, to bound $\|x^{(\mathcal{K})} - x^{(0)}\|$, it suffices bound $\|x^{(\mathcal{K})} - y^{(\mathcal{K}-1)}\|$, which satisfies

$$\|x^{(\mathcal{K})} - y^{(\mathcal{K}-1)}\| = \eta\|\hat{H}^{-1}\nabla f(y^{(\mathcal{K}-1)})\|,$$

where

$$\begin{aligned} &\hat{H}^{-1}\nabla f(y^{(\mathcal{K}-1)}) \\ &= \hat{H}^{-1}\nabla f(y^{(\mathcal{K}-2)}) + \hat{H}^{-1} \int_{y=y^{(\mathcal{K}-2)}}^{y^{(\mathcal{K}-1)}} \nabla^2 f(y) dy \\ &= \hat{H}^{-1}\nabla f(y^{(\mathcal{K}-2)}) + \hat{H}^{-1}H(y^{(\mathcal{K}-1)} - y^{(\mathcal{K}-2)}) + \int_{y=y^{(\mathcal{K}-2)}}^{y^{(\mathcal{K}-1)}} \hat{H}^{-1}(H - \nabla^2 f(y)) dy. \quad (23) \end{aligned}$$

The first and the second term satisfy

$$\|\hat{H}^{-1}\nabla f(y^{(\mathcal{K}-2)})\| = \|x^{(\mathcal{K}-1)} - y^{(\mathcal{K}-2)}\|/\eta \leq 3B/\eta,$$

and

$$\|\hat{H}^{-1}H(y^{(\mathcal{K}-2)} - y^{(\mathcal{K}-1)})\| \leq \|y^{(\mathcal{K}-2)} - y^{(\mathcal{K}-1)}\| \leq 4B,$$

respectively. As for the third term, given that $\|\hat{H}^{-1}\| \leq (12\delta p_{\max})^{-1}$ by Corollary 3 and

$$\|\hat{H} - \nabla^2 f(y)\| \leq \delta + 4L_2B \leq 2\delta,$$

it follows that

$$\left\| \int_{y=y^{(\mathcal{K}-2)}}^{y^{(\mathcal{K}-1)}} \hat{H}^{-1}(H - \nabla^2 f(y)) dy \right\| \leq B.$$

Therefore,

$$\|x^{(\mathcal{K})} - y^{(\mathcal{K}-1)}\| \leq (3 + 4\eta + 1)B = 5B,$$

and we can conclude that

$$\|x^{(\mathcal{K})} - x^{(0)}\| \leq \|y^{(\mathcal{K}-1)} - x^{(0)}\| + \|x^{(\mathcal{K})} - y^{(\mathcal{K}-1)}\| \leq 7B.$$

□

E.3 Function Value Decrease Case

In this subsection, we discuss the decrease in the function value of Algorithm 1 in the case where the “if condition” in Line 7 is triggered.

Given that $\overline{H}$ is symmetric, we can find a set of orthonormal basis $\{\hat{h}_1, \dots, \hat{h}_d\}$ such that each $\hat{h}_i$ is an eigenvector of $\overline{H}$ with eigenvalue λ_i . We decompose these coordinates into two sets

$$\mathcal{S}_{\text{sc}} := \left\{ i : \lambda_i \geq -\frac{\theta}{\eta} \right\}, \quad \mathcal{S}_{\text{nc}} := \left\{ i : \lambda_i < -\frac{\theta}{\eta} \right\}, \quad (24)$$

where sc and nc abbreviate strongly convex and not strongly convex, respectively. We further define the corresponding projectors

$$\Pi_{\text{sc}} := \sum_{i \in \mathcal{S}_{\text{sc}}} \hat{h}_i \hat{h}_i^\top, \quad \Pi_{\text{nc}} := \sum_{i \in \mathcal{S}_{\text{nc}}} \hat{h}_i \hat{h}_i^\top. \quad (25)$$

For any $\hat{v} \in \mathbb{R}^d$, denote

$$\hat{v}_{\text{sc}} := \Pi_{\text{sc}} \hat{v}, \quad \hat{v}_{\text{nc}} := \Pi_{\text{nc}} \hat{v}$$

and

$$\begin{aligned} \hat{g}_{\text{sc}}(\hat{v}) &:= \langle \nabla \hat{f}(\hat{x}^{(0)}), \hat{v}_{\text{sc}} - \hat{x}_{\text{sc}}^{(0)} \rangle + \frac{1}{2} (\hat{v}_{\text{sc}} - \hat{x}_{\text{sc}}^{(0)})^\top \overline{H} (\hat{v}_{\text{sc}} - \hat{x}_{\text{sc}}^{(0)}) \\ \hat{g}_{\text{nc}}(\hat{v}) &:= \langle \nabla \hat{f}(\hat{x}^{(0)}), \hat{v}_{\text{nc}} - \hat{x}_{\text{nc}}^{(0)} \rangle + \frac{1}{2} (\hat{v}_{\text{nc}} - \hat{x}_{\text{nc}}^{(0)})^\top \overline{H} (\hat{v}_{\text{nc}} - \hat{x}_{\text{nc}}^{(0)}) \end{aligned} \quad (26)$$

Then, we have $\hat{g}(\hat{v}) = \hat{g}_{\text{sc}}(\hat{v}) + \hat{g}_{\text{nc}}(\hat{v})$.

E.3.1 Function Value Decrease of $\hat{g}_{\text{sc}}$

The proof structure in this part is similar to the proof of [29, Lemma 2].

Lemma 21. *Let $\eta \leq 1/4$. Then for any $0 \leq k \leq \mathcal{K} - 1$ and any $\alpha > 0$, we have*

$$\begin{aligned} \hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}^{(k)}) &\leq -\frac{1}{2} (\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H} (\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) + \frac{\|\ell^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{1}{2\eta} \left(\|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \left(1 - \frac{\alpha\eta}{12\delta p_{\max}} \right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \right). \end{aligned}$$

Proof. Given that $\hat{g}_{\text{sc}}$ is quadratic, for any two consecutive iterations, we have

$$\begin{aligned} \hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) &= \hat{g}_{\text{sc}}(\hat{x}^{(k)}) + \langle \nabla \hat{g}_{\text{sc}}(\hat{x}^{(k)}), \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle \\ &\quad + \frac{1}{2} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)})^\top \overline{H} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}), \end{aligned}$$

where

$$\begin{aligned} \nabla \hat{g}_{\text{sc}}(\hat{x}^{(k)}) &= \nabla \hat{f}_{\text{sc}}(\hat{y}^{(k)}) + (\nabla \hat{g}_{\text{sc}}(\hat{y}^{(k)}) - \nabla \hat{f}_{\text{sc}}(\hat{y}^{(k)})) + (\nabla \hat{g}_{\text{sc}}(\hat{x}^{(k)}) - \nabla \hat{g}_{\text{sc}}(\hat{y}_{\text{sc}}^{(k)})) \\ &= -\frac{1}{\eta} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}) + \overline{H} (\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) - \hat{\ell}_{\text{sc}}^{(k)}. \end{aligned}$$

Hence,

$$\begin{aligned} &\hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}_{\text{sc}}^{(k)}) \\ &= -\frac{1}{\eta} \langle \hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}, \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle - \langle \hat{\ell}_{\text{sc}}^{(k)}, \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle \\ &\quad + (\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}) + \frac{1}{2} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)})^\top \overline{H} (\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}), \end{aligned}$$

where

$$\langle \hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}, \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle = \|x_{\text{sc}}^{(k)} - y_{\text{sc}}^{(k)}\|^2 - \|x_{\text{sc}}^{(k+1)} - y_{\text{sc}}^{(k)}\|^2 - \|x_{\text{sc}}^{(k+1)} - x_{\text{sc}}^{(k)}\|^2$$

and

$$\begin{aligned} & (\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}) + \frac{1}{2}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}) \\ &= \frac{1}{2}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}) - \frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}). \end{aligned}$$

Furthermore, for any $\alpha > 0$ we have

$$\begin{aligned} -\langle \hat{l}_{\text{sc}}^{(k)}, \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle &= -\langle \hat{l}^{(k)}, \hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)} \rangle \\ &\leq \frac{1}{2\alpha} \|\hat{H}^{1/2} \hat{l}\|^2 + \frac{\alpha}{2} \|\hat{H}^{-1/2}(\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)})\|^2 \\ &\leq \frac{\|\hat{l}^{(k)}\|^2}{2\alpha} + \frac{\alpha}{24\delta p_{\max}} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \end{aligned}$$

by Corollary 3. It then follows that

$$\begin{aligned} & \hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}^{(k)}) \\ &= \frac{1}{2}(\hat{x}_{\text{sc}}^{(k+1)} - y_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k+1)} - y_{\text{sc}}^{(k)}) - \frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - y_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - y_{\text{sc}}^{(k)}) + \frac{\|\hat{l}^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{1}{2\eta} \left(\|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \left(1 - \frac{\alpha\eta}{12\delta p_{\max}}\right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \right) \\ &\leq -\frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) + \frac{\|\hat{l}^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{1}{2\eta} \left(\|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \left(1 - \frac{\alpha\eta}{12\delta p_{\max}}\right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \right), \end{aligned}$$

where the last inequality follows from

$$\frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - y_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - y_{\text{sc}}^{(k)}) - \frac{1}{2\eta} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{y}_{\text{sc}}^{(k)}\|^2 \leq 0.$$

since $\overline{H} \preceq I$ and $\eta \leq 1/4$. We can therefore conclude that

$$\begin{aligned} \hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}^{(k)}) &\leq -\frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) + \frac{\|\hat{l}^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{1}{2\eta} \left(\|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \left(1 - \frac{\alpha\eta}{12\delta p_{\max}}\right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \right). \end{aligned}$$

□

Lemma 22. Let $\eta \leq 1/4$ and $0 < \theta \leq 1$. In the case where the “if condition” in Line 7 of Algorithm 1 is triggered, we have

$$\hat{g}_{\text{sc}}(\hat{x}^{(\mathcal{K})}) - \hat{g}_{\text{sc}}(\hat{x}^{(0)}) \leq -\frac{\theta}{4\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 + \frac{\eta L_2^2 B^4 \mathcal{K}}{3\theta \delta p_{\max}}.$$

Proof. By Lemma 21, for any $\alpha > 0$ the following inequality holds:

$$\begin{aligned} \hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}^{(k)}) &\leq -\frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) + \frac{\|\hat{l}^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{1}{2\eta} \left(\|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2 - \left(1 - \frac{\alpha\eta}{12\delta p_{\max}}\right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 \right), \end{aligned}$$

where

$$-\frac{1}{2}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}) \leq \frac{\theta}{2\eta} \|\hat{x}_{\text{sc}}^{(k)} - \hat{y}_{\text{sc}}^{(k)}\|^2$$

as per the definition of $\mathcal{S}_{\text{sc}}$ in (24) and Π_{sc} in (25), which leads to

$$\begin{aligned}\hat{g}_{\text{sc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{sc}}(\hat{x}^{(k)}) &\leq -\frac{1}{2\eta} \left(1 - \frac{\alpha\eta}{12\delta p_{\max}}\right) \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 + \frac{\|\ell^{(k)}\|^2}{2\alpha} \\ &\quad + \frac{(1+\theta)(1-\theta)^2}{2\eta} \|\hat{x}_{\text{sc}}^{(k)} - \hat{x}_{\text{sc}}^{(k-1)}\|^2,\end{aligned}$$

given that $\hat{y}_{\text{sc}}^{(k)} - \hat{x}_{\text{sc}}^{(k)} = (1-\theta)(\hat{x}_{\text{sc}}^{(k)} - \hat{x}_{\text{sc}}^{(k-1)})$ as per (7). Define the potential function

$$\xi_{\text{sc}}^{(k)} := \hat{g}_{\text{sc}}(\hat{x}^{(k)}) + \frac{(1+\theta)(1-\theta)^2}{2\eta} \|\hat{x}_{\text{sc}}^{(k)} - \hat{x}_{\text{sc}}^{(k-1)}\|^2$$

and set $\alpha = 6\delta p_{\max}\theta/\eta$. Then, we have

$$\xi_{\text{sc}}^{(k+1)} - \xi_{\text{sc}}^{(k)} \leq -\frac{\theta}{4\eta} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 + \frac{\eta\|\ell^{(k)}\|^2}{12\delta p_{\max}\theta}.$$

Summing over all the iterations in this epoch, we can conclude that

$$\begin{aligned}\hat{g}_{\text{sc}}(\hat{x}^{(\mathcal{K})}) - \hat{g}_{\text{sc}}(\hat{x}^{(0)}) &\leq -\frac{\theta}{4\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 + \frac{\eta}{12\delta p_{\max}\theta} \sum_{k=0}^{\mathcal{K}-1} \|\ell^{(k)}\|^2 \\ &\leq -\frac{\theta}{4\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}_{\text{sc}}^{(k+1)} - \hat{x}_{\text{sc}}^{(k)}\|^2 + \frac{\eta L_2^2 B^4 \mathcal{K}}{3\delta p_{\max}\theta}.\end{aligned}$$

□

E.3.2 Function Value Decrease of $\hat{g}_{\text{nc}}$

Lemma 23. Let $\eta \leq 1/4$ and $0 < \theta \leq 1$. In the case where the “if condition” in Line 7 of Algorithm 1 is triggered, we have

$$\hat{g}_{\text{nc}}(\hat{x}^{(\mathcal{K})}) - \hat{g}_{\text{nc}}(\hat{x}^{(0)}) \leq -\frac{\theta}{2\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}\|^2 + \frac{\eta L_2^2 B^4 \mathcal{K}}{6\theta\delta p_{\max}}.$$

Proof. The proof of this lemma follows a similar structure as the proof of [29, Lemma 3]. Denote

$\hat{u} := \hat{x}_{\text{nc}}^{(0)} - \bar{H}^\top \nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)})$, which allows us to rewrite $\hat{g}_{\text{nc}}(\hat{v})$ as

$$\hat{g}_{\text{nc}}(\hat{v}) = \frac{1}{2}(\hat{v}_{\text{nc}} - \hat{u})^\top \bar{H}(\hat{v}_{\text{nc}} - \hat{u}) - \frac{1}{2}(\nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)}))^\top \bar{H}^{-1} \nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)}), \quad \forall \hat{v} \in \mathbb{R}^d.$$

Then for any $0 \leq k \leq \mathcal{K} - 1$, we have

$$\begin{aligned}\hat{g}_{\text{nc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{nc}}(\hat{x}^{(k)}) &= \frac{1}{2}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{u})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{u}) - \frac{1}{2}(\hat{x}_{\text{nc}}^{(k)} - \hat{u})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ &= \frac{1}{2}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k+1)} + 2\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ &= \frac{1}{2}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}) + (\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}),\end{aligned}$$

where the first term is upper bounded by

$$\frac{1}{2}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \bar{H}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}) \leq -\frac{\theta}{2\eta} \|\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}\|^2$$

following the definition of $\mathcal{S}_{\text{nc}}$ in (24) and Π_{nc} in (25). As for the second term, note that

$$\begin{aligned}\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)} &= \hat{y}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k)} - \eta \nabla \hat{g}_{\text{nc}}(\hat{y}^{(k)}) - \eta \hat{\ell}_{\text{nc}}^{(k)} \\ &= (1-\theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)}) - \eta \bar{H}(\hat{y}_{\text{nc}}^{(k)} - \hat{u}) - \eta \hat{\ell}_{\text{nc}}^{(k)} \\ &= (1-\theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)}) - \eta \bar{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u} + (1-\theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})) - \eta \hat{\ell}_{\text{nc}}^{(k)}.\end{aligned}$$

where the second line follows from the observation that

$$\begin{aligned}\overline{H}(\hat{y}_{\text{nc}}^{(k)} - \hat{u}) &= \overline{H}(\hat{y}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(0)}) + \overline{H}(\hat{x}_{\text{nc}}^{(0)} - \hat{u}) \\ &= \nabla \hat{g}_{\text{nc}}(\hat{y}^{(k)}) - \nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)}) + \overline{H}\overline{H}^\dagger \nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)}) = \nabla \hat{g}_{\text{nc}}(\hat{y}^{(k)}).\end{aligned}$$

Hence,

$$\begin{aligned}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ = (1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) - \eta \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u})\|^2 \\ - \eta(1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}^2(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) - \eta \langle \hat{\iota}_{\text{nc}}^{(k)}, \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \rangle,\end{aligned}\quad (27)$$

where we have

$$\begin{aligned}-\eta(1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}^2(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ \leq \frac{\eta(1 - \theta)}{2} \left(\|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})\|^2 + \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u})\|^2 \right)\end{aligned}$$

and

$$-\eta \langle \hat{\iota}_{\text{nc}}^{(k)}, \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \rangle \leq \frac{\eta}{2(1 + \theta)} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2 + \frac{\eta(1 + \theta)}{2} \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u})\|^2.$$

Combined with (27), we obtain

$$\begin{aligned}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ \leq (1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ + \frac{\eta(1 - \theta)}{2} \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})\|^2 + \frac{\eta}{2(1 + \theta)} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2 \\ = (1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k-1)} - \hat{u}) + (1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)}) \\ + \frac{\eta(1 - \theta)}{2} \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})\|^2 + \frac{\eta}{2(1 + \theta)} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2 \\ \leq (1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k-1)} - \hat{u}) + \frac{\eta}{2(1 + \theta)} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2,\end{aligned}$$

where the last inequality follows from the fact that

$$(1 - \theta)(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)}) + \frac{\eta(1 - \theta)}{2} \|\overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})\|^2 \geq 0$$

since $\eta \leq 1/4$ and $\|\overline{H}\| \leq 1$. Hence,

$$\begin{aligned}(\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k)} - \hat{u}) \\ \leq (1 - \theta)^k (\hat{x}_{\text{nc}}^{(k)} - \hat{x}_{\text{nc}}^{(k-1)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(k-1)} - \hat{u}) + \frac{\eta}{2(1 + \theta)} \sum_{t=1}^k (1 - \theta)^{k-t} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2 \\ \leq \frac{\eta}{2} \sum_{t=1}^k (1 - \theta)^{k-t} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2,\end{aligned}$$

where the last inequality follows from

$$(\hat{x}_{\text{nc}}^{(1)} - \hat{x}_{\text{nc}}^{(0)})^\top \overline{H}(\hat{x}_{\text{nc}}^{(0)} - \hat{u}) = (-\nabla \hat{g}_{\text{nc}}(\hat{x}_0)) \cdot \overline{H} \cdot (\overline{H}^\dagger \nabla \hat{g}_{\text{nc}}(\hat{x}^{(0)})) \leq 0.$$

Furthermore, for each $0 \leq k \leq \mathcal{K} - 1$, by Lemma 19 and Corollary 3 we have

$$\|\hat{\iota}_{\text{nc}}^{(k)}\|^2 \leq \|\hat{\iota}^{(k)}\|^2 = \|\hat{H}^{-1/2} \hat{H}^{1/2} \hat{\iota}^{(k)}\|^2 \leq \|\hat{H}^{-1/2}\|^2 \cdot \|\hat{\iota}^{(k)}\|^2 \leq \frac{L_2^2 B^4}{3\delta p_{\max}},$$

which leads to

$$\frac{\eta}{2} \sum_{t=1}^k (1 - \theta)^{k-t} \|\hat{\iota}_{\text{nc}}^{(k)}\|^2 \leq \frac{\eta L_{\text{nc}}^2 B^4}{6\theta \delta p_{\max}}$$

and

$$\hat{g}_{\text{nc}}(\hat{x}^{(k+1)}) - \hat{g}_{\text{nc}}(\hat{x}^{(k)}) \leq -\frac{\theta}{2\eta} \|\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}\|^2 + \frac{\eta L_2^2 B^4}{6\theta\delta p_{\max}}.$$

We can thus conclude that

$$\hat{g}_{\text{nc}}(\hat{x}^{(\mathcal{K})}) - \hat{g}_{\text{nc}}(\hat{x}^{(0)}) \leq -\frac{\theta}{2\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}_{\text{nc}}^{(k+1)} - \hat{x}_{\text{nc}}^{(k)}\|^2 + \frac{\eta L_2^2 B^4 \mathcal{K}}{6\theta\delta p_{\max}}.$$

□

E.3.3 Function Value Decrease of f

Proposition 4. Let $\eta \leq 1/4$ and $0 < \theta \leq 1$. In the case where the “if condition” in Line 7 of Algorithm 1 is triggered, we have

$$f(x^{(\mathcal{K})}) - f(x^{(0)}) \leq -\frac{3\theta\delta B^2 \lceil \log_2(L_1/\delta) \rceil}{K\eta} + 36L_2 B^3 \leq -\sqrt{\frac{\tilde{\epsilon}^3}{L_2}},$$

where $\tilde{\epsilon} = \epsilon/p_{\max}^8$ is defined in (4).

Proof. Combining Lemma 22 and Lemma 23, we obtain

$$\hat{g}(\hat{x}^{(\mathcal{K})}) - \hat{g}(\hat{x}^{(0)}) \leq -\frac{\theta}{4\eta} \sum_{k=0}^{\mathcal{K}-1} \|\hat{x}^{(k+1)} - \hat{x}^{(k)}\|^2 + \frac{2\eta L_2^2 B^4 \mathcal{K}}{3\theta\delta p_{\max}} \leq -\frac{3\theta\delta p_{\max} B^2}{\eta \mathcal{K}} + \frac{2\eta L_2^2 B^4 \mathcal{K}}{3\theta\delta p_{\max}},$$

following the condition in Line 7. Then, we can conclude that

$$\begin{aligned} f(x^{(\mathcal{K})}) - f(x^{(0)}) &= g(x^{(\mathcal{K})}) - g(x^{(0)}) + (f(x^{(\mathcal{K})}) - g(x^{(\mathcal{K})})) - (f(x^{(0)}) - g(x^{(0)})) \\ &\leq -\frac{3\theta\delta p_{\max} B^2}{\eta \mathcal{K}} + \frac{2\eta L_2^2 B^4 \mathcal{K}}{3\theta\delta p_{\max}} + 60L_2 B^3 \leq -\sqrt{\frac{\tilde{\epsilon}^3}{L_2}} \end{aligned}$$

given that f is L_2 -Hessian Lipschitz, and $\|x^{(\mathcal{K})} - x^{(0)}\| \leq 7B$ by Lemma 20. □

E.4 Small Gradient Case

In this subsection, we provide an upper bound on the gradient of the output of Algorithm 1 in the case where the last iterate stays close enough to $x^{(0)}$, or more concretely, the “if condition” in Line 7 is not triggered. Similar to Section E.3, we use $\{\hat{h}_1, \dots, \hat{h}_d\}$ to denote the set of orthonormal vectors such that $\hat{h}_i$ is the eigenvector of $\bar{H}$ with eigenvalue λ_i . We decompose these coordinates into $\Theta(p_{\max})$ sets:

$$\mathcal{S}_p := \{i : \bar{l}_p < \lambda_i \leq \bar{l}_{p+1}\}, \quad \forall p \in \mathbb{N}^+ \text{ and } p_{\min} \leq p \leq p_{\max}, \quad (28)$$

where the definition of $\bar{l}_p$ is given in (14). Then, the projector onto the eigenspace of eigenvectors with indices in $\mathcal{S}_p$ equals

$$\sum_{i \in \mathcal{S}_p} \hat{h}_i \hat{h}_i^\top = \bar{\Pi}_p, \quad \forall p \in \mathbb{N}^+ \text{ and } p_{\min} \leq p \leq p_{\max} \quad (29)$$

where $\bar{\Pi}_p$ is defined in (18). Moreover, define

$$\mathcal{S}_{\text{base}} := \{i : \lambda_i < \bar{l}_{p_{\min}}\}. \quad (30)$$

Then, the projector onto the eigenspace of eigenvectors with indices in $\mathcal{S}_p$ equals

$$\sum_{i \in \mathcal{S}_{\text{base}}} \hat{h}_i \hat{h}_i^\top = \bar{\Pi}_{\text{base}}, \quad (31)$$

where $\bar{\Pi}_{\text{base}}$ is defined in (19). For any $\hat{v} \in \mathbb{R}^d$, denote

$$\hat{v}_{\text{base}} := \bar{\Pi}_{\text{base}} \hat{v}, \quad \text{and} \quad \hat{v}_p := \bar{\Pi}_p \hat{v}, \quad \forall p \in \mathbb{N}^+ \text{ and } p_{\min} \leq p \leq p_{\max}.$$

Moreover, for any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$, we define

$$\hat{g}_p(\hat{v}) := \langle \nabla \hat{f}(\hat{x}^{(0)}), \hat{v}_p - \hat{x}_p^{(0)} \rangle + \frac{1}{2} (\hat{v}_p - \hat{x}_p^{(0)})^\top \bar{H} (\hat{v}_p - \hat{x}_p^{(0)}), \quad (32)$$

and

$$\hat{g}_{\text{base}}(\hat{v}) := \langle \nabla \hat{f}(\hat{x}^{(0)}), \hat{v}_{\text{base}} - \hat{x}_{\text{base}}^{(0)} \rangle + \frac{1}{2} (\hat{v}_{\text{base}} - \hat{x}_{\text{base}}^{(0)})^\top \bar{H} (\hat{v}_{\text{base}} - \hat{x}_{\text{base}}^{(0)}), \quad (33)$$

Then, we have $\hat{g}(\hat{v}) = \hat{g}_{\text{base}}(\hat{v}) + \sum_{p=p_{\min}}^{p_{\max}} \hat{g}_p(\hat{v})$ and

$$\nabla \hat{g}(\hat{v}) = \nabla \hat{g}_{\text{base}}(\hat{v}) + \sum_{p=p_{\min}}^{p_{\max}} \nabla \hat{g}_p(\hat{v}).$$

E.4.1 Small Gradient of $\hat{g}_p$ at x^{out}

In this part, we show that the gradient of $\hat{g}_p(x^{\text{out}})$ is small for any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$. We define

$$M := I - \eta \bar{H}, \quad M_p := \Pi_p \cdot M \cdot \Pi_p.$$

Lemma 24. For any $v^{(k)} \in \mathbb{R}^d$ with the initial condition $v^{(-1)} = v^{(0)}$ and $\gamma^{(-1)} = 0$, that satisfies the following recursion formula

$$v^{(k+1)} = a \cdot M v^{(k)} - b \cdot M v^{(k-1)} + \hat{\ell}^{(k)},$$

for some symmetric matrix $M \in \mathbb{R}^{d \times d}$ and $a, b \in \mathbb{R}$, we have

$$v^{(k)} = \psi_k(M) v^{(0)} + \sum_{j=0}^{k-1} \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \gamma^{(j)},$$

where

$$\psi_k(M) := \frac{I - \mathcal{Q}(M)}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \mathcal{P}(M)^{k+1} + \frac{\mathcal{P}(M) - I}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \mathcal{Q}(M)^{k+1}, \quad \forall k \geq 0, \quad (34)$$

and

$$\mathcal{P}(M) := \frac{aM + \sqrt{a^2 M^2 - 4bM}}{2}, \quad \mathcal{Q}(M) := \frac{aM - \sqrt{a^2 M^2 - 4bM}}{2}. \quad (35)$$

Proof. The solution to the homogeneous part $v^{(k+1)} = a \cdot M v^{(k)} - b \cdot M v^{(k-1)}$ is

$$v^{(k)} = \psi_k(M) v^{(0)},$$

where

$$\psi_k(M) := \frac{I - \mathcal{Q}(M)}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \mathcal{P}(M)^{k-1} + \frac{\mathcal{P}(M) - I}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \mathcal{Q}(M)^{k-1}$$

with

$$\mathcal{P}(M) = \frac{aM + \sqrt{a^2 M^2 - 4bM}}{2}, \quad \mathcal{Q}(M) = \frac{aM - \sqrt{a^2 M^2 - 4bM}}{2}.$$

Counting in the inhomogeneous part, for each $\hat{\ell}^{(j)}$, it leads to the following additional term in $v^{(k)}$ for any $k \geq j$:

$$\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \gamma^{(j)}.$$

We can conclude that

$$v^{(k)} = \psi_k(M) v^{(0)} + \sum_{j=0}^{k-1} \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \gamma^{(j)}.$$

□

Lemma 25 (Properties of $\mathcal{P}(M_p)$ and $\mathcal{Q}(M_p)$). Let $a := 2 - \theta$, $b := 1 - \theta$. If $\eta \leq 1/4$ and $\theta \leq \eta/(2p_{\max})$, for any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$, the matrices $\mathcal{P}(M_p)$ and $\mathcal{Q}(M_p)$ defined in (35) satisfy

1. $\|\mathcal{P}(M_p)\| = \|\mathcal{Q}(M_p)\| = b\|M_p\| \leq 1 - \eta/p_{\max}$;
2. $\left\| \frac{1}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \leq \sqrt{\frac{p_{\max}}{\eta}}$;
3. $\left\| \frac{I - \mathcal{Q}(M_p)}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| = \left\| \frac{\mathcal{P}(M_p) - I}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \leq 2\sqrt{\frac{p_{\max}}{\eta}}$

Proof. Observe that

$$\mathcal{P}(M_p) := \frac{aM_p + \sqrt{a^2M_p^2 - 4bM_p}}{2}, \quad \mathcal{Q}(M_p) := \frac{aM_p - \sqrt{a^2M_p^2 - 4bM_p}}{2}.$$

Given the definition of Π_p in (16), M_p satisfies

$$(1 - \eta r_p)I \preceq M_p \preceq (1 - \eta l_p)I. \quad (36)$$

If $\eta \leq 1/4$ and $\theta \leq \eta/(2p_{\max})$, we have $a^2M_p^2 - 4bM_p \preceq 0$, which leads to

$$\mathcal{P}(M_p) := \frac{aM_p + i\sqrt{-a^2M_p^2 + 4bM_p}}{2}, \quad \mathcal{Q}(M_p) := \frac{aM_p - i\sqrt{-a^2M_p^2 + 4bM_p}}{2}$$

Therefore, we have

$$\begin{aligned} \|\mathcal{P}(M_p)\| = \|\mathcal{Q}(M_p)\| &= \frac{1}{2} \|(aM_p)^2 + (-a^2M_p^2 + 4bM_p)\|^{1/2} \\ &\leq \sqrt{b\|M_p\|} \leq 1 - \eta/p_{\max}. \end{aligned}$$

As for the second entry, since

$$\mathcal{P}(M_p) - \mathcal{Q}(M_p) = i\sqrt{-a^2M_p^2 + 4bM_p},$$

by (36) and the value of a, b we have

$$\left\| \frac{1}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \leq \sqrt{\frac{p_{\max}}{\eta}}$$

which leads to

$$\begin{aligned} \left\| \frac{I - \mathcal{Q}(M_p)}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| &= \left\| \frac{\mathcal{P}(M_p) - I}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \\ &\leq (1 + \|\mathcal{P}(M_p)\|) \cdot \left\| \frac{1}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \leq 2\sqrt{\frac{p_{\max}}{\eta}}. \end{aligned}$$

□

Lemma 26 (Bound on the difference between $\nabla \hat{f}_p$ and $\nabla \hat{g}_p$). For any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$, in the case where the “if condition” in Line 7 is not triggered, for any iteration k of Algorithm 1, we have $\|\hat{\iota}_p^{(k)}\| \leq 2^{6-p/2}p_{\max}\|\iota^{(k)}\|/\sqrt{\delta}$.

Proof. It follows from Proposition 2 that

$$\|\hat{\iota}_p^{(k)}\| = \|\bar{\Pi}_p \hat{\iota}\| \leq \frac{2}{\sqrt{\phi(2^p \delta)}} \|\hat{H}^{1/2} \bar{\Pi}_p \hat{\iota}^{(k)}\|,$$

where

$$\|\hat{H}^{1/2} \hat{\iota}\| \geq \frac{2^{-5}}{1 + 2^{-p/4}p_{\max}} \cdot \|\hat{H}^{1/2} \bar{\Pi}_p \hat{\iota}^{(k)}\| \geq \frac{\|\hat{H}^{1/2} \bar{\Pi}_p \hat{\iota}^{(k)}\|}{64p_{\max}}$$

by Proposition 3, which leads to

$$\|\hat{\iota}_p^{(k)}\| \leq \frac{2^{6-p/2}p_{\max}}{\sqrt{\delta}} \|\hat{H}^{1/2} \hat{\iota}\| = \frac{2^{6-p/2}p_{\max}}{\sqrt{\delta}} \|\iota^{(k)}\|.$$

□

Lemma 27. If $\eta \leq \frac{1}{4}$ and $\theta \leq \eta/(2p_{\max})$, then for any $p \in \mathbb{N}^+$ with $p_{\min} \leq p \leq p_{\max}$ and any $0 < j \leq k-1$ we have

$$\left\| \widehat{H}^{1/2} \overline{H} \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} \right\| \leq \frac{128p_{\max}^{5/2}}{\sqrt{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k-j} \|\iota^{(j)}\|.$$

Proof. It follows from Lemma 25 that

$$\begin{aligned} \left\| \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} \right\| &\leq \left\| \frac{1}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \cdot \|\mathcal{P}(M_p)^{k-j}\| \cdot \|\hat{\iota}_p^{(j)}\| \\ &\quad + \left\| \frac{1}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \cdot \|\mathcal{Q}(M_p)^{k-j}\| \cdot \|\hat{\iota}_p^{(j)}\| \\ &\leq 2\sqrt{\frac{p_{\max}}{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k-j} \|\hat{\iota}_p^{(j)}\| \end{aligned}$$

where

$$\|\hat{\iota}_p^{(j)}\| \leq \frac{2^{6-p/2} p_{\max}}{\sqrt{\delta}} \|\iota^{(j)}\|.$$

by Lemma 26. Since

$$\overline{H} \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} = \overline{\Pi}_p \overline{H} \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)},$$

we have

$$\begin{aligned} &\left\| \widehat{H}^{1/2} \overline{H} \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} \right\| \\ &= \left\| \widehat{H}^{1/2} \overline{\Pi}_p \overline{H} \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} \right\| \\ &\leq 2^{p/2} p_{\max} \sqrt{\delta} \cdot \left\| \overline{\Pi}_p \left(\frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \right) \hat{\iota}_p^{(j)} \right\| \\ &\leq 2^{1+p/2} p_{\max}^{3/2} \sqrt{\frac{\delta}{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k-j} \|\hat{\iota}_p^{(j)}\| \\ &\leq \frac{128p_{\max}^{5/2}}{\sqrt{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k-j} \|\iota^{(j)}\|. \end{aligned}$$

where the first inequality is by Proposition 2, and the second inequality follows from the fact that $\|\overline{H}\| \leq 1$. $\square$

Lemma 28. If $\eta \leq \frac{1}{4}$ and $\theta \leq \eta/(2p_{\max})$, then for any $p \in \mathbb{N}^+$ with $p_{\min} \leq p \leq p_{\max}$, any vector $\hat{v} \in \mathbb{R}^d$, and any integer $k \geq 0$, we have

$$\|\widehat{H}^{1/2} \overline{H} \psi_j(M_p) \hat{v}_p\| \leq 2^{2+p/2} p_{\max}^{3/2} \sqrt{\frac{\delta}{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k+1} \|\hat{v}_p\|.$$

Proof. By the definition of ψ in (34) we have

$$\psi_k(M_p) \hat{v}_p = \frac{I - \mathcal{Q}(M_p)}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \cdot \mathcal{P}(M_p)^{k+1} \hat{v}_p + \frac{\mathcal{P}(M_p) - I}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \cdot \mathcal{Q}(M_p)^{k+1} \hat{v}_p$$

and

$$\begin{aligned} \|\psi_k(M_p) \hat{v}_p\| &\leq \left\| \frac{I - \mathcal{Q}(M_p)}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \cdot \|\mathcal{P}(M_p)^{k+1}\| \cdot \|\hat{v}_p\| \\ &\quad + \left\| \frac{\mathcal{P}(M_p) - I}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \right\| \cdot \|\mathcal{Q}(M_p)^{k+1}\| \cdot \|\hat{v}_p\| \\ &\leq 4\sqrt{\frac{p_{\max}}{\eta}} \left(1 - \frac{\eta}{p_{\max}} \right)^{k+1} \|\hat{v}_p\| \end{aligned}$$

by Lemma 25. Since $\bar{H}\psi_k(M_p)\hat{v}_p = \bar{\Pi}_p\bar{H}\psi_k(M_p)\hat{v}_p$, we have

$$\begin{aligned}\|\hat{H}^{1/2}\bar{H}\psi_k(M_p)\hat{v}_p\| &= \|\hat{H}^{1/2}\bar{\Pi}_p\bar{H}\psi_k(M_p)\hat{v}_p\| \\ &\leq 2^{p/2}p_{\max}\sqrt{\delta} \cdot \|\bar{\Pi}_p\bar{H}\psi_k(M_p)\hat{v}_p\| \\ &\leq 2^{p/2}p_{\max}\sqrt{\delta}\|\psi_k(M_p)\hat{v}_p\| \\ &\leq 2^{2+p/2}p_{\max}^{3/2}\sqrt{\frac{\delta}{\eta}}\left(1 - \frac{\eta}{p_{\max}}\right)^{k+1}\|\hat{v}_p\|.\end{aligned}$$

where the first inequality is by Proposition 2, and the second inequality follows from the fact that $\|\bar{H}\| \leq 1$. $\square$

Lemma 29. *If $\eta \leq 1/4$ and $\theta \leq \eta/(2p_{\max})$, for any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$, in the case where the “if condition” in Line 7 is not triggered, for any iteration $k \geq 0$ of Algorithm 1 we have*

$$\|\hat{H}^{1/2}\nabla\hat{g}_p(\hat{x}^{(k)})\| \leq \frac{2^{4+p/2}\delta p_{\max}^2 B}{\eta^{3/2}}\left(1 - \frac{\eta}{p_{\max}}\right)^{k+1} + \frac{256L_2B^2p_{\max}^{7/2}}{\eta^{1/2}}$$

Proof. For any iteration k , by (7) we have

$$\begin{aligned}\hat{x}_p^{(k+1)} &= \hat{y}_p^{(k)} - \eta\nabla\hat{f}_p(\hat{y}_p^{(k)}) \\ &= \hat{y}_p^{(k)} - \eta\bar{H}(\hat{y}_p^{(k)} - \hat{x}_p^{(0)}) - \eta\nabla\hat{f}_p(\hat{x}_p^{(0)}) - \eta\hat{t}_p^{(k)} \\ &= \hat{x}_p^{(k)} - \eta\nabla\hat{f}_p(\hat{x}_p^{(0)}) + (1 - \theta)(\hat{x}_p^{(k)} - \hat{x}_p^{(k-1)}) \\ &\quad - \eta\bar{H}(\hat{x}_p^{(k)} - \hat{x}_p^{(0)} + (1 - \theta)(\hat{x}_p^{(k)} - \hat{x}_p^{(k-1)})) - \eta\hat{t}_p^{(k)}.\end{aligned}$$

Denote $\tilde{x}_p^{(k)} := \hat{x}_p^{(k)} - \hat{x}_p^{(0)} + \bar{H}^\dagger\nabla\hat{f}_p(\hat{x}_p^{(0)})$. Then, the above equation is equivalent to

$$\begin{aligned}\tilde{x}_p^{(k+1)} &= \tilde{x}_p^{(k)} + (1 - \theta)(\tilde{x}_p^{(k)} - \tilde{x}_p^{(k-1)}) - \eta\bar{H}(\tilde{x}_p^{(k)} + (1 - \theta)(\tilde{x}_p^{(k)} - \tilde{x}_p^{(k-1)})) - \eta\hat{t}_p^{(k)} \\ &= (I - \eta\bar{H})((2 - \theta)\tilde{x}_p^{(k)} - (1 - \theta)\tilde{x}_p^{(k-1)}) - \eta\hat{t}_p^{(k)}.\end{aligned}$$

Let $a := 2 - \theta$, $b := 1 - \theta$. By Lemma 24 we have

$$\tilde{x}_p^{(k)} = \psi_k(M_p) \cdot \tilde{x}_p^{(0)} - \eta \sum_{j=0}^{k-1} \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \hat{t}_p^{(j)}, \quad (37)$$

where

$$\psi_k(M_p) := \frac{I - \mathcal{Q}(M_p)}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \cdot \mathcal{P}(M_p)^{k+1} + \frac{\mathcal{P}(M_p) - I}{\mathcal{P}(M_p) - \mathcal{Q}(M_p)} \cdot \mathcal{Q}(M_p)^{k+1},$$

which leads to

$$\begin{aligned}\hat{H}^{1/2}\nabla\hat{g}_p(\hat{x}^{(k)}) &= \hat{H}^{1/2}\bar{H}\tilde{x}_p^{(k)} = \hat{H}^{1/2}\bar{H}\psi_k(M_p)\tilde{x}_p^{(0)} \\ &\quad - \eta \sum_{j=0}^{k-1} \hat{H}^{1/2}\bar{H} \cdot \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \hat{t}_p^{(j)}.\end{aligned} \quad (38)$$

For the first term in (38), by Lemma 28 we have

$$\|\hat{H}^{1/2}\bar{H}\psi_k(M_p)\tilde{x}_p^{(0)}\| \leq 2^{2+p/2}p_{\max}^{3/2}\sqrt{\frac{\delta}{\eta}}\left(1 - \frac{\eta}{p_{\max}}\right)^{k+1}\|\tilde{x}_p^{(0)}\|$$

where

$$\|\tilde{x}_p^{(0)}\| = \|\bar{H}^\dagger\nabla\hat{f}_p(\hat{x}_p^{(0)})\| \leq \sqrt{p_{\max}}\|\nabla\hat{f}_p(\hat{x}_p^{(0)})\|$$

and

$$\|\nabla\hat{f}_p(\hat{x}_p^{(0)})\| \leq \|\nabla\hat{f}(\hat{x}^{(0)})\| = \frac{\|\hat{x}^{(1)} - \hat{x}^{(0)}\|}{\eta} = \frac{4\sqrt{\delta p_{\max}}B}{\eta}$$

which leads to

$$\|\hat{H}^{1/2}\bar{H}\psi_k(M_p)\tilde{x}_p^{(0)}\| \leq \frac{2^{4+p/2}\delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{k+1}.$$

As for the second term of (38), by Lemma 27 we have

$$\begin{aligned} \left\| \eta \sum_{j=0}^{k-1} \hat{H}^{1/2}\bar{H} \cdot \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \hat{\iota}_p^{(j)} \right\| &\leq \eta \sum_{j=0}^{k-1} \left\| \hat{H}^{1/2}\bar{H} \cdot \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \hat{\iota}_p^{(j)} \right\| \\ &\leq 128 p_{\max}^{5/2} \eta^{1/2} \sum_{j=0}^{k-1} \left(1 - \frac{\eta}{p_{\max}}\right)^{k-j} \|\iota^{(j)}\|, \end{aligned}$$

where for each j we have

$$\|\iota^{(j)}\| \leq \frac{1}{2} L_2 \|y^{(j)} - x^{(0)}\|^2 \leq 2 L_2 B^2$$

given that f is L_2 -Hessian Lipschitz. Hence,

$$\begin{aligned} \left\| \eta \sum_{j=0}^{k-1} \hat{H}^{1/2}\bar{H} \cdot \frac{\mathcal{P}(M)^{k-j} - \mathcal{Q}(M)^{k-j}}{\mathcal{P}(M) - \mathcal{Q}(M)} \cdot \hat{\iota}_p^{(j)} \right\| \\ \leq 256 L_2 B^2 p_{\max}^{5/2} \eta^{1/2} \sum_{j=0}^{k-1} \left(1 - \frac{\eta}{p_{\max}}\right)^{k-j} \leq 256 L_2 B^2 p_{\max}^{7/2} / \eta^{1/2}. \end{aligned}$$

We can therefore conclude that

$$\|\hat{H}^{1/2}\nabla\hat{g}_p(\hat{x}^{(k)})\| \leq \frac{2^{4+p/2}\delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{k+1} + \frac{256 L_2 B^2 p_{\max}^{7/2}}{\eta^{1/2}}.$$

□

Lemma 30. If $\eta \leq 1/4$ and $\theta \leq \eta/(2p_{\max})$, for any $p \in \mathbb{N}^+$ and $p_{\min} \leq p \leq p_{\max}$, in the case where the “if condition” in Line 7 is not triggered, we have

$$\|\hat{H}^{1/2}\nabla\hat{g}_p(\hat{x}^{\text{out}})\| \leq \frac{2^{6+p/2}\delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{2^{10} L_2 B^2 p_{\max}^{7/2}}{\eta^{1/2}}$$

Proof. Given that $\hat{y}^{(k)} = \hat{x}^{(k)} + (1 - \theta)(\hat{x}^{(k)} - \hat{x}^{(k-1)})$ and that $\hat{g}_p$ is quadratic, for any $k \geq 1$ we have

$$\nabla\hat{g}_p(\hat{y}^{(k)}) = \nabla\hat{g}_p(\hat{x}^{(k)}) + (1 - \theta)(\nabla\hat{g}_p(\hat{x}^{(k)}) - \nabla\hat{g}_p(\hat{x}^{(k-1)})),$$

which leads to

$$\begin{aligned} \|\hat{H}^{1/2}\nabla\hat{g}_p(\hat{y}^{(k)})\| &\leq (2 - \theta)\|\hat{H}^{1/2}\hat{g}_p(\hat{x}^{(k)})\| + (1 - \theta)\|\hat{H}^{1/2}\hat{g}_p(\hat{x}^{(k-1)})\| \\ &\leq 2\|\hat{H}^{1/2}\hat{g}_p(\hat{x}^{(k)})\| + \|\hat{H}^{1/2}\hat{g}_p(\hat{x}^{(k-1)})\| \\ &\leq \frac{2^{6+p/2}\delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^k + \frac{2^{10} L_2 B^2 p_{\max}^{7/2}}{\eta^{1/2}} \end{aligned}$$

by Lemma 29. Furthermore, since

$$\hat{x}^{\text{out}} = \frac{1}{K_0 + 1 - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \hat{y}^{(t)},$$

we have

$$\begin{aligned} \|\hat{H}^{1/2}\nabla\hat{g}_p(\hat{x}^{\text{out}})\| &\leq \frac{1}{K_0 + 1 - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \|\hat{H}^{1/2}\hat{g}_p(\hat{y}^{(t)})\| \\ &\leq \frac{2^{6+p/2}\delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{2^{10} L_2 B^2 p_{\max}^{7/2}}{\eta^{1/2}}. \end{aligned}$$

□

E.4.2 Small Gradient of $\hat{g}_{\text{base}}$ at x^{out}

Lemma 31. *If $\eta \leq 1/4$ and $\theta \leq \eta/2$, in the case where the “if condition” in Line 7 is not triggered, we have*

$$\|\hat{H}^{1/2} \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})\| \leq \frac{32\ell\sqrt{\delta p_{\max}}B}{\eta K^2} + \frac{8\theta\ell\sqrt{\delta p_{\max}}B}{\eta K} + 80p_{\max}^{3/2}L_2B^2$$

for any ℓ that satisfies

$$\|\hat{H}^{1/2}\hat{v}_{\text{base}}\| \leq \ell \cdot \|\hat{v}_{\text{base}}\|, \quad \forall \hat{v} \in \mathbb{R}^d. \quad (39)$$

Proof. The proof of this lemma has a similar structure as the proof of [29, Lemma 5]. Given that $\hat{g}_{\text{base}}$ is quadratic, $\nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})$ can be expressed as

$$\nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}}) = \frac{1}{K_0 + 1 - K/2} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \nabla \hat{g}_{\text{base}}(\hat{y}^{(k)}),$$

where we have

$$\begin{aligned} -\nabla \hat{g}_{\text{base}}(\hat{y}^{(k)}) &= \frac{1}{\eta} (\hat{x}_{\text{base}}^{(k+1)} - \hat{y}_{\text{base}}^{(k)}) + \hat{t}_{\text{base}} \\ &= \frac{1}{\eta} \left((\hat{x}_{\text{base}}^{(k+1)} - \hat{x}_{\text{base}}^{(k)}) - (1 - \theta)(\hat{x}_{\text{base}}^{(k)} - \hat{x}_{\text{base}}^{(k-1)}) \right) + \hat{t}_{\text{base}} \end{aligned}$$

and

$$-\eta(K_0 + 1 - K/2) \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}}) = \hat{x}_{\text{base}}^{(K_0+1)} - \hat{x}_{\text{base}}^{(K_0)} + \theta(\hat{x}_{\text{base}}^{(K_0)} - \hat{x}_{\text{base}}^{(K/2)}) + \eta \sum_{k=\lfloor K/2 \rfloor}^{K_0} \hat{t}_{\text{base}},$$

where the last term satisfies

$$\begin{aligned} \left\| \hat{H}^{1/2} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \hat{t}_{\text{base}} \right\| &\leq \sum_{k=\lfloor K/2 \rfloor}^{K_0} \|\hat{H}^{1/2} \hat{t}_{\text{base}}\| \leq 40p_{\max}^{3/2} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \|\hat{H}^{1/2} \hat{t}\| \\ &\leq 80(K_0 + 1 - K/2)p_{\max}^{3/2}L_2B^2 \end{aligned}$$

by Lemma 18. Hence,

$$\begin{aligned} &\|\hat{H}^{1/2} \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})\| \\ &\leq \frac{4}{\eta K} \|\hat{H}^{1/2}(\hat{x}_{\text{base}}^{(K_0+1)} - \hat{x}_{\text{base}}^{(K_0)})\| + \frac{4\theta}{\eta K} \|\hat{H}^{1/2}(\hat{x}_{\text{base}}^{(K_0)} - \hat{x}_{\text{base}}^{(K/2)})\| + 80p_{\max}^{3/2}L_2B^2 \\ &\leq \frac{4\ell}{\eta K} \|\hat{x}_{\text{base}}^{(K_0+1)} - \hat{x}_{\text{base}}^{(K_0)}\| + \frac{4\theta\ell}{\eta K} \|\hat{x}_{\text{base}}^{(K_0)} - \hat{x}_{\text{base}}^{(K/2)}\| + 80p_{\max}^{3/2}L_2B^2 \\ &\leq \frac{4\ell}{\eta K} \|\hat{x}_{\text{base}}^{(K_0+1)} - \hat{x}_{\text{base}}^{(K_0)}\| + \frac{4\theta\ell}{\eta K} \|\hat{x}_{\text{base}}^{(K_0)} - \hat{x}_{\text{base}}^{(K/2)}\| + 80p_{\max}^{3/2}L_2B^2 \end{aligned}$$

where the second inequality is due to the condition given in (39). By the condition $K_0 \leftarrow \text{argmin}_{\lfloor \frac{3K}{4} \rfloor \leq t \leq K-1} \|\hat{x}^{(t+1)} - \hat{x}^{(t)}\|$, we have

$$\|\hat{x}_{\text{base}}^{(K_0+1)} - \hat{x}_{\text{base}}^{(K_0)}\|^2 \leq \frac{1}{K - \lfloor 3K/4 \rfloor} \sum_{k=\lfloor 3K/4 \rfloor}^{K-1} \|x_{\text{base}}^{(k+1)} - x_{\text{base}}^{(k)}\|^2 \leq \frac{48\delta p_{\max}B^2}{K^2}$$

given that the “if condition” in Line 7 is not triggered, which leads to

$$\|\hat{H}^{1/2} \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})\| \leq \frac{32\ell\sqrt{\delta p_{\max}}B}{\eta K^2} + \frac{8\theta\ell\sqrt{\delta p_{\max}}B}{\eta K} + 80p_{\max}^{3/2}L_2B^2$$

□

Corollary 5. *If $\eta \leq 1/4$ and $\theta \leq \eta/2$, in the case where the “if condition” in Line 7 is not triggered, we have*

$$\|\hat{H}^{1/2} \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})\| \leq \frac{32p_{\max}^{3/2}\delta B}{\eta K^2} + \frac{8\theta p_{\max}^{3/2}\delta B}{\eta K} + 80p_{\max}^{3/2}L_2B^2$$

Proof. The desired inequality follows by combining Lemma 31 with Corollary 4. □

E.4.3 Small Gradient of f at x^{out}

Proposition 5. *With the choice of parameters in Theorem 2, in the case where the “if condition” in Line 7 is not triggered, we have $\|\nabla f(x^{\text{out}})\| \leq \epsilon$.*

Proof. By Lemma 30 and Corollary 5, we have

$$\begin{aligned} \|\nabla g(x^{\text{out}})\| &= \|\hat{H}^{1/2} \hat{g}(\hat{x}^{\text{out}})\| = \left\| \hat{H}^{1/2} \left(\nabla \hat{g}(\hat{x}^{\text{out}}) + \sum_{p=p_{\min}}^{p_{\max}} \nabla \hat{g}_p(\hat{x}^{\text{out}}) \right) \right\| \\ &\leq \|\hat{H}^{1/2} \nabla \hat{g}_{\text{base}}(\hat{x}^{\text{out}})\| + \sum_{p=p_{\min}}^{p_{\max}} \|\hat{H}^{1/2} \nabla \hat{g}_p(\hat{x}^{\text{out}})\| \\ &\leq \frac{2^{6+p/2} \delta p_{\max}^3 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{32 p_{\max}^{3/2} \delta B}{\eta K^2} + \frac{8 \theta p_{\max}^{3/2} \delta B}{\eta K} + \frac{2^{10} L_2 B^2 p_{\max}^{9/2}}{\eta^{1/2}} \\ &= \frac{2^{6+p/2} \delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{40 p_{\max}^{3/2} \delta B}{\eta K^2} + \frac{2^{10} L_2 B^2 p_{\max}^{9/2}}{\eta^{1/2}}, \end{aligned}$$

and

$$\begin{aligned} \|\nabla f(x^{\text{out}})\| &\leq \|\nabla g(x^{\text{out}})\| + \|\nabla f(x^{\text{out}}) - \nabla g(x^{\text{out}})\| \\ &\leq \|\nabla g(x^{\text{out}})\| + 2L_2 B^2 \\ &\leq \frac{2^{6+p/2} \delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{40 p_{\max}^{3/2} \delta B}{\eta K^2} + \frac{2^{11} L_2 B^2 p_{\max}^{9/2}}{\eta^{1/2}}. \end{aligned} \quad (40)$$

Given that

$$K \geq \frac{2p_{\max}}{\eta} \log \left(\frac{3 \times 2^{6+p/2} \delta p_{\max}^2 B}{\eta^{3/2} \epsilon} \right),$$

the first term of (40) satisfies

$$\frac{2^{6+p/2} \delta p_{\max}^2 B}{\eta^{3/2}} \left(1 - \frac{\eta}{p_{\max}}\right)^{K/2} + \frac{40 p_{\max}^{3/2} \delta B}{\eta K^2} + \frac{2^{11} L_2 B^2 p_{\max}^{9/2}}{\eta^{1/2}} \leq \frac{\epsilon}{3}.$$

Furthermore, the second and the third term satisfy

$$\frac{40 p_{\max}^{3/2} \delta B}{\eta K^2} \leq \frac{\epsilon}{3}, \quad \frac{2^{11} L_2 B^2 p_{\max}^{9/2}}{\eta^{1/2}} \leq \frac{\epsilon}{3},$$

respectively. We can thus conclude that $\|\nabla f(x^{\text{out}})\| \leq \epsilon$. $\square$

E.5 Putting Everything Together

In this section, we give the proof of Theorem 2, and present an additional Lemma that characterizes the suboptimality of x^{out} , the output of Algorithm 1.

Proof of Theorem 2. By combining Proposition 4 and Proposition 5, we know that at least one of the two conditions in the theorem statement must hold. As for the distance between x^{out} and $x^{(0)}$, in the case where the “if condition” in Line 7 is triggered, we have

$$\|x^{\text{out}} - x^{(0)}\| = \|x^{(\mathcal{K})} - x^{(0)}\| \leq 7B$$

by Lemma 20. Otherwise,

$$\|x^{\text{out}} - x^{(0)}\| \leq \frac{1}{K_0 + 1 - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \|y^{(k)} - x^{(0)}\| \leq 2B$$

by Lemma 19. $\square$

Lemma 32. *In the case where the “if condition” in Line 7 is triggered, the output x^{out} of Algorithm 1 satisfies*

$$f(x^{\text{out}}) - f(x^{(0)}) \leq \frac{6\delta\epsilon}{L_2} + \frac{L_2^{1/2}\epsilon^{3/2}}{162}.$$

Proof. In the case where the “if condition” in Line 7 is triggered, we have

$$x^{\text{out}} \leftarrow \frac{1}{K_0 + 1 - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} y^{(k)},$$

where in each iteration k we have

$$\|\widehat{H}^{1/2}(y^{(k)} - x^{(0)})\| \leq 2\|\widehat{H}^{1/2}(x^{(k)} - x^{(0)})\| + \|\widehat{H}^{1/2}(x^{(k-1)} - x^{(0)})\| \leq 6\sqrt{3\delta p_{\max}}B,$$

which leads to

$$\|\widehat{H}^{1/2}(x^{\text{out}} - x^{(0)})\| \leq \frac{1}{K_0 + 1 - \lfloor K/2 \rfloor} \sum_{k=\lfloor K/2 \rfloor}^{K_0} \|\widehat{H}^{1/2}(y^{(k)} - x^{(0)})\| \leq 6\sqrt{3\delta p_{\max}}B.$$

Hence,

$$\begin{aligned} g(x^{\text{out}}) - g(x^{(0)}) &= \frac{1}{2}(x^{\text{out}} - x^{(0)})^\top \nabla^2 f(x^{(0)})(x^{\text{out}} - x^{(0)}) \\ &\leq \frac{1}{2}(x^{\text{out}} - x^{(0)})^\top \widehat{H}(x^{\text{out}} - x^{(0)}) \leq 54\delta p_{\max}B^2 \leq \frac{6\delta\epsilon}{L_2}. \end{aligned}$$

Meanwhile,

$$f(x^{\text{out}}) - g(x^{\text{out}}) \leq \frac{L_2}{6} \|x^{\text{out}} - x^{(0)}\|^3 \leq \frac{\epsilon^{3/2}}{6L_2^{1/2}},$$

by which we can conclude that

$$f(x^{\text{out}}) - f(x^{(0)}) \leq \frac{6\delta\epsilon}{L_2} + \frac{\epsilon^{3/2}}{6L_2^{1/2}}.$$

□

F Analysis of Algorithm 2

Lemma 1. *Suppose $\epsilon \leq L_1^2/L_2$. In each iteration t of Algorithm 2 before it terminates, we have*

$$f(x^{(t+1)}) - f(x^{(t)}) \leq -\tilde{p}^{-12} \sqrt{\epsilon^3/L_2}$$

if $x^{(t+1)}$ is not ϵ -critical for f , where we denote $\tilde{p} = \max\{\lceil \log(L_1/\tilde{\delta}) \rceil, 16\}$.

Proof. If the current H satisfies $H \prec -2\tilde{\delta}I$, by Line 7 and Line 8 we have

$$\begin{aligned} f(x^{(t+1)}) - f(x^{(t)}) &\leq \langle \nabla f(x^{(t)}), Rv \rangle + \frac{R^2}{2} v^\top \nabla^2 f(x^{(t)})v + \frac{L_2 R^3}{6} \\ &\leq \frac{R^2}{2} v^\top \nabla^2 f(x^{(t)})v + \frac{\tilde{\delta} R^2}{6}, \end{aligned}$$

where

$$\begin{aligned} v^\top \nabla^2 f(x^{(t)})v &= v^\top H v + v^\top (\nabla^2 f(x^{(t)}) - H)v \\ &\leq -2\tilde{\delta} + \|\nabla^2 f(x^{(t)}) - H\| \\ &\leq -2\tilde{\delta} + \min\{\|\nabla^2 f(x^{(t)}) - \nabla^2 f(\bar{x})\| + \|\nabla^2 f(\bar{x}) - H\|, 2L_1\} \\ &= -2\tilde{\delta} + \min\{L_2 R + \delta, 2L_1\} = -\tilde{\delta}, \end{aligned}$$

which leads to

$$f(x^{(t+1)}) - f(x^{(t)}) \leq -\frac{1}{3}\tilde{\delta}R^2 \leq -\frac{1}{3}L_2R^3 \leq -\frac{1}{\tilde{p}^{12}}\sqrt{\frac{\epsilon^3}{L_2}}.$$

Otherwise,

$$x^{(t+1)} = \text{Critical-or-Progress}(x^{(t)}, H, 2\tilde{\delta}, \epsilon, L_1, L_2),$$

which by Theorem 2 satisfies

$$f(x^{(t+1)}) - f(x^{(t)}) \leq -\frac{1}{\tilde{p}^{12}}\sqrt{\frac{\epsilon^3}{L_2}}$$

if $\|\nabla f(x^{(t+1)})\| > \epsilon$. □

Lemma 33. Let $0 < \epsilon \leq \min\{L_1^2L_2^{-1}, \Delta^{2/3}L_2^{1/3}\}$. The output x^{out} of Algorithm 2 satisfies

$$\|x^{\text{out}} - x^{(0)}\| \leq \frac{3\Delta}{\epsilon} \log^8 \left(\frac{L_1}{c_\delta} + 16 \right)$$

and

$$f(x^{\text{out}}) - f(x^{(0)}) \leq \frac{54c_\delta\epsilon}{L_2} \log^8 \left(\frac{L_1}{c_\delta} + 16 \right) + \frac{\epsilon^{3/2}}{6L_2^{1/2}},$$

where $c_\delta := \min\{L_1, \delta + \Delta L_2/(n_H\epsilon)\}$.

Proof. Suppose Algorithm 2 terminates at the $\mathcal{T}$ -th iteration. For any $t < \mathcal{T} - 1$, by Proposition 4 we have

$$f(x^{(t+1)}) - f(x^{(t)}) \leq \frac{1}{\tilde{p}^{12}}\sqrt{\frac{\epsilon^3}{L_2}},$$

indicating that

$$\mathcal{T} \leq \tilde{p}^{12}\Delta\sqrt{L_2/\epsilon^3}.$$

Since

$$\|x^{(t+1)} - x^{(t)}\| \leq \frac{7}{3\tilde{p}^4}\sqrt{\frac{\epsilon}{L_2}}$$

in each iteration t by (4) and Theorem 2, we have

$$\|x^{\text{out}} - x^{(0)}\| = \|x^{(\mathcal{T})} - x^{(0)}\| \leq 3\tilde{p}^8\Delta/\epsilon \leq \frac{3\Delta}{\epsilon} \log^8 \left(\frac{L_1}{c_\delta} + 16 \right)$$

As for the function value change in last iteration, by Lemma 32, we have

$$f(x^{\text{out}}) = f(x^{(\mathcal{T})}) \leq f(x^{(\mathcal{T}-1)}) + \frac{54\tilde{\delta}\epsilon}{L_2} + \frac{L_2\epsilon^{3/2}}{6L_2^{1/2}}.$$

Summing over all the iterations, we can conclude that

$$f(x^{\text{out}}) - f(x^{(0)}) \leq \frac{18c_\delta\epsilon}{L_2} \log^8 \left(\frac{L_1}{c_\delta} + 16 \right) + \frac{\epsilon^{3/2}}{6L_2^{1/2}}.$$

□

G Analysis of Algorithm 3

Lemma 34. Given $f := \mathbb{R}^d \rightarrow \mathbb{R}$ with L_2 -Lipschitz Hessian, for any $x \in \mathbb{R}^d$ and any symmetric H satisfying $\|H - \nabla^2 f(x)\| \leq \delta$, denote

$$y \leftarrow x - (\Pi_{\text{large}} H \Pi_{\text{large}})^\dagger \nabla f(x).$$

If $\ell \geq \max\{24\Delta_x^{1/3} L_2^{2/3}, 2\delta\}$ for $\Delta_x = f(x) - \inf_{z \in \mathbb{R}^d} f(z)$, we have

$$\|\Pi_{\text{large}} \nabla f(y)\| \leq 2\delta \sqrt{\frac{3\Delta_x}{\ell}} + \frac{6L_2 \Delta_x}{\ell}$$

and

$$\|\Pi_{\text{small}} \nabla f(y)\| \leq \|\Pi_{\text{small}} \nabla f(x)\| + 2\delta \sqrt{\frac{3\Delta_x}{\ell}} + \frac{6L_2 \Delta_x}{\ell}.$$

Proof. Denote $u := y - x = -(\Pi_{\text{large}} H \Pi_{\text{large}})^\dagger \nabla f(x)$. We first show that $\|u\| \leq \ell/L_2$. Assume the contrary, we have

$$f\left(x + \frac{\ell}{L_2} \cdot \frac{u}{\|u\|}\right) - f(x) \leq -\frac{\ell}{2} \left(\frac{\ell}{L_2}\right)^2 + \frac{\delta}{2} \left(\frac{\ell}{L_2}\right)^2 + \frac{L_2}{6} \left(\frac{\ell}{L_2}\right)^3 \leq -\frac{\ell^3}{12L_2^2} \leq -2\Delta_x,$$

contradiction. Then, we have

$$-\Delta_x \leq f(y) - f(x) \leq -\frac{1}{2} u^\top \nabla^2 f(x) u + \frac{1}{6} L_2 \|u\|^3 \leq \frac{\ell \|u\|^2}{12}$$

which leads to $\|u\| \leq 2\sqrt{3\Delta_x/\ell}$. Then, we have

$$\begin{aligned} \|\Pi_{\text{large}} \nabla f(y)\| &\leq \|\Pi_{\text{large}} \nabla f(x) - \Pi_{\text{large}} H u\| + \|\Pi_{\text{large}} (H - \nabla^2 f(x)) u\| + \frac{1}{2} L_2 \|u\|^2 \\ &\leq \delta \|u\| + \frac{1}{2} L_2 \|u\|^2 \leq 2\delta \sqrt{\frac{3\Delta_x}{\ell}} + \frac{6L_2 \Delta_x}{\ell}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \|\Pi_{\text{small}} \nabla f(y)\| &\leq \|\Pi_{\text{small}} \nabla f(x)\| + \|\Pi_{\text{small}} H u\| + \|\Pi_{\text{small}} (H - \nabla^2 f(x)) u\| + \frac{1}{2} L_2 \|u\|^2 \\ &\leq \|\Pi_{\text{small}} \nabla f(x)\| + 2\delta \sqrt{\frac{3\Delta_x}{\ell}} + \frac{6L_2 \Delta_x}{\ell}. \end{aligned}$$

□

Theorem 4. Let $\text{Alg}(f_{\leq L_1}, L_1, L_2, \delta, \Delta, \epsilon)$ be a procedure that, for any function $f_{\leq L_1}: \mathbb{R}^d \rightarrow \mathbb{R}$ with L_1 -Lipschitz gradient, L_2 -Lipschitz Hessian, and Δ -bounded suboptimality, uses

- n_H queries to a δ -approximate Hessian oracle for $f_{\leq L_1}$, and,
- $n_g(L_1, L_2, \delta, \Delta, \epsilon)$ queries to a gradient oracle for $f_{\leq L_1}$,

and returns an $\epsilon/2$ -critical point x^{out} satisfying $\|x^{\text{out}} - x^{(0)}\| \leq R_{\text{out}}$ and $f(x^{\text{out}}) - \inf_z f(z) \leq \Delta_{\text{out}}$. Then, for any f with L_2 -Lipschitz Hessian and Δ -bounded suboptimality, any $0 < \epsilon \leq \min\{L_1^2 L_2^{-1}, \Delta^{2/3} L_2^{1/3}\}$, and any ℓ that satisfies

$$\ell \geq \max\left\{\frac{800\Delta}{\epsilon^2} (L_2 R_{\text{out}} + \delta)^2, \frac{48L_2 \Delta_{\text{out}}}{\epsilon}, 24\Delta_{\text{out}}^{1/3} L_2^{2/3}, 2\delta\right\}. \quad (12)$$

Algorithm 3 returns an ϵ -critical point using

- $n_H + 1$ queries to a δ -approximate Hessian oracle for f , and,
- $n_g(\ell, L_2, \delta, \Delta, \epsilon)$ queries to a gradient oracle for f .

Proof. By Lemma 34 we have

$$\begin{aligned}\|\nabla f(y)\| &\leq \|\Pi_{\text{large}} \nabla f(y)\| + \|\Pi_{\text{small}} \nabla f(y)\| \\ &\leq \|\Pi_{\text{small}} \nabla f(x^{\text{out}})\| + 4\|H - \nabla^2 f(x^{\text{out}})\| \sqrt{\frac{3\Delta_{\text{out}}}{\ell}} + \frac{12L_2\Delta_{\text{out}}}{\ell} \\ &\leq \frac{\epsilon}{2} + 4(\delta + L_2R_{\text{out}}) \sqrt{\frac{3\Delta_{\text{out}}}{\ell}} + \frac{12L_2\Delta_{\text{out}}}{\ell} \leq \epsilon\end{aligned}$$

given the choice of ℓ in (12). $\square$

Corollary 2. Let $f: \mathbb{R}^d \mapsto \mathbb{R}$ L_2 -Lipschitz Hessian. Given any $x^{(0)} \in \mathbb{R}^d$ with Δ -bounded suboptimality with respect to f , any positive integer $n_H \geq 1$, and $0 < \epsilon \leq \Delta^{2/3}L_2^{1/3}$, Algorithm 3 using Algorithm 2 as the subroutine Alg outputs an ϵ -critical point of f with at most n_H queries to a δ -approximate Hessian oracle and

$$O\left(\frac{\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\delta + \frac{\Delta L_2}{n_H \epsilon}} \cdot \text{poly log} \left(\frac{1}{c_\delta} \left(\frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta^2}{\epsilon^2} + \delta\right)\right)\right)$$

queries to a gradient oracle for f .

Proof. Set

$$\begin{aligned}\Delta_{\text{out}} &= 54 \left(\Delta + \frac{\delta \epsilon}{L_2}\right) \log^9 \left(\frac{\hat{\ell}}{c_\delta} + 16\right) + \frac{\epsilon^{3/2}}{6L_2^{1/2}}, \\ \ell &:= \hat{\ell} \log^{19}(\hat{\ell}/c_\delta), \\ R_{\text{out}} &= \frac{3\Delta}{\epsilon} \log^9 \left(\frac{\hat{\ell}}{c_\delta} + 16\right),\end{aligned}\tag{41}$$

where

$$\hat{\ell} := \max \left\{ \frac{800\Delta}{\epsilon^2} \left(\frac{3L_2\Delta}{\epsilon} + \delta\right)^2, 2\delta \right\} = O\left(\frac{L_2^2\Delta^3}{\epsilon^4} + \frac{\Delta\delta^2}{\epsilon^2} + \delta\right).\tag{42}$$

Observe that the above parameters satisfy

$$\ell \geq \max \left\{ \frac{800\Delta}{\epsilon^2} (L_2 R_{\text{out}} + \delta)^2, \frac{48L_2\Delta_{\text{out}}}{\epsilon}, 24\Delta_{\text{out}}^{1/3} L_2^{2/3}, 2\delta \right\},$$

which gives

$$\|x^{\text{out}} - x^{(0)}\| \leq R_{\text{out}}, \quad f(x^{\text{out}}) - \inf_{x \in \mathbb{R}^d} f(x) \leq \Delta_{\text{out}}$$

by Lemma 33, where x^{out} is the output of Restarted-Approx-Hessian-AGD when applied to $f_{\leq \ell}$. Then by Theorem 4, Algorithm 3 outputs an ϵ -critical point. Since Restarted-Approx-Hessian-AGD starts by querying the δ -approximate Hessian oracle at $x^{(0)}$, the query in Line 3 can be reused, and there are a total of at most n_H queries to a δ -approximate Hessian oracle and

$$\begin{aligned}&\frac{2\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\min \left\{ \ell, \delta + \frac{\Delta L_2}{n_H \epsilon} \right\}} \cdot \log^{18} \left(\frac{\hat{\ell}}{c_\delta} + 16\right) \\ &= O\left(\frac{\Delta L_2^{1/4}}{\epsilon^{7/4}} \sqrt{\delta + \frac{\Delta L_2}{n_H \epsilon}} \cdot \text{poly log} \left(\frac{1}{c_\delta} \left(\frac{L_2^2 \Delta^3}{\epsilon^4} + \frac{\Delta \delta^2}{\epsilon^2} + \delta\right)\right)\right)\end{aligned}$$

queries to a gradient oracle. $\square$