Mathematical Programming
https://doi.org/10.1007/s10107-022-01913-5

FULL LENGTH PAPER

Series A

Hessian averaging in stochastic Newton methods achieves superlinear convergence

Sen Na[1] · Michał Dereziński[2] · Michael W. Mahoney[1]

Received: 23 April 2022 / Accepted: 26 November 2022

Abstract

We consider minimizing a smooth and strongly convex objective function using a stochastic Newton method. At each iteration, the algorithm is given an oracle access to a stochastic estimate of the Hessian matrix. The oracle model includes popular algorithms such as Subsampled Newton and Newton Sketch, which can efficiently construct stochastic Hessian estimates for many tasks, e.g., training machine learning models. Despite using second-order information, these existing methods do not exhibit superlinear convergence, unless the stochastic noise is gradually reduced to zero during the iteration, which would lead to a computational blow-up in the per-iteration cost. We propose to address this limitation with Hessian averaging: instead of using the most recent Hessian estimate, our algorithm maintains an average of all the past estimates. This reduces the stochastic noise while avoiding the computational blow-up. We show that this scheme exhibits local Q-superlinear convergence with a non-asymptotic rate of $(\Upsilon\sqrt{\log(t)/t})^t$, where Υ is proportional to the level of stochastic noise in the Hessian oracle. A potential drawback of this (uniform averaging) approach is that the averaged estimates contain Hessian information from the global phase of the method, i.e., before the iterates converge to a local neighborhood. This leads to a distortion that may substantially delay the superlinear convergence until long after the local neighborhood is reached. To address this drawback, we study a number of weighted averaging schemes that assign larger weights to recent Hessians, so that the superlinear convergence arises sooner, albeit with a slightly slower rate. Remarkably, we show that there exists a universal weighted averaging scheme that transitions to

✉ Sen Na
senna@berkeley.edu

Michał Dereziński
derezin@umich.edu

Michael W. Mahoney
mmahoney@stat.berkeley.edu

1 ICSI and Department of Statistics, University of California, Berkeley, USA

2 Computer Science and Engineering, University of Michigan, Ann Arbor, USA

Published online: 13 December 2022

local convergence at an optimal stage, and still exhibits a superlinear convergence rate nearly (up to a logarithmic factor) matching that of uniform Hessian averaging.

Keywords Stochastic Newton method · Hessian averaging · Superlinear convergence

Mathematics Subject Classification 90-08 · 90-10 · 90C06 · 90C15 · 90C25

1 Introduction

We consider minimizing a smooth and strongly convex objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}), \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is twice continuously differentiable with $\mathbf{H}(\mathbf{x})=\nabla^2 f(\mathbf{x})\in\mathbb{R}^{d\times d}$ being its Hessian matrix. We suppose the Hessian $\mathbf{H}(\mathbf{x})$ satisfies

$$\lambda_{\min}\cdot\mathbf{I}\preceq\mathbf{H}(\mathbf{x})\preceq\lambda_{\max}\cdot\mathbf{I},\quad\forall\mathbf{x}\in\mathbb{R}^d,$$

for some constants $0<\lambda_{\min}\le\lambda_{\max}$, and we let $\kappa=\lambda_{\max}/\lambda_{\min}$ denote its condition number. We also let $\mathbf{x}^\star=\arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ be the unique global solution.

Problem (1) is arguably the most basic optimization problem, which nevertheless arises in many applications in machine learning and statistics [7, 9, 35, 56]. There is a plethora of methods for solving (1), and each (type of) method has its own benefits under favorable settings. This paper particularly focuses on solving (1) via second-order methods, where a (approximated) Hessian matrix is involved in the scheme. Consider the classical Newton's method of the form

$$\mathbf{x}_{t+1}=\mathbf{x}_t-\mu_t\mathbf{H}_t^{-1}\nabla f_t, \tag{2}$$

where $\nabla f_t=\nabla f(\mathbf{x}_t)$, $\mathbf{H}_t=\mathbf{H}(\mathbf{x}_t)$, and μ_t is selected by passing a line search condition. Classical results indicate that Newton's method in (2) exploits a global Q-linear convergence in the error of function value $f(\mathbf{x}_t)-f(\mathbf{x}^\star)$; and a local Q-quadratic convergence in the iterate error $\|\mathbf{x}_t-\mathbf{x}^\star\|$. More precisely, Newton's method has two phases, separated by a neighborhood

$$\mathcal{N}_\nu=\{\mathbf{x}:\|\mathbf{x}-\mathbf{x}^\star\|\le\nu\},\qquad\text{for some }\nu>0.$$

When $\mathbf{x}_t\notin\mathcal{N}_\nu$, (2) is in the *damped Newton phase*, where the objective $f(\mathbf{x}_t)$ is decreased by at least a fixed amount in each iteration, and converges linearly. In this phase, $\|\mathbf{x}_t-\mathbf{x}^\star\|$ may converge slowly (e.g., it provably converges R-linearly using the fact that $\|\mathbf{x}_t-\mathbf{x}^\star\|^2\le 2/\lambda_{\min}(f(\mathbf{x}_t)-f(\mathbf{x}^\star))$). When $\mathbf{x}_t\in\mathcal{N}_\nu$, (2) transits to the *quadratically convergent phase*, where the unit stepsize is accepted and $\|\mathbf{x}_t-\mathbf{x}^\star\|$ converges quadratically. See [8, Section 9.5] for the analysis. Compared to first-order methods, although second-order methods often exploit a faster convergence rate and

behave more robustly to tuning parameters, they hinge on a high computational cost of forming the exact Hessian matrix $\mathbf{H}_t$ at each step. To resolve such a computational bottleneck, which is particularly pressing in large-scale data applications, a variety of deterministic and stochastic methods have been proposed for constructing different alternatives of the Hessian matrix.

When a deterministic alternative of $\mathbf{H}_t$, say $\check{\mathbf{H}}_t$, is employed in (2), which may come from a finite difference approximation of the second-order derivative, or from a quasi-Newton update such as BFGS or DFP, the convergence behavior is well understood. Specifically, if $\{\check{\mathbf{H}}_t\}_t$ are positive definite with uniform upper and lower bounds, the damped phase is preserved by the same analysis as Newton's method. Furthermore, the quadratically convergent phase is weakened to a superlinearly convergent phase, and the superlinear convergence occurs if and only if the celebrated Dennis-Moré condition [14] holds, i.e.,

$$\lim_{t\to\infty} \frac{\|(\check{\mathbf{H}}_t - \mathbf{H}_t)\check{\mathbf{H}}_t^{-1}\nabla f_t\|}{\|\check{\mathbf{H}}_t^{-1}\nabla f_t\|} = 0. \tag{3}$$

See [47, Theorem 3.7] for the analysis. Recently, a deeper understanding of quasi-Newton methods for minimizing smooth and strongly convex objectives has been reported in [31, 50–52]. These works enhanced the analyses based on Dennis-Moré condition in (3) by performing a *local, non-asymptotic* convergence analysis, and provided explicit superlinear rates for different potential functions of quasi-Newton methods. The non-asymptotic superlinear results are more informative than asymptotic superlinear results established via (3), i.e., $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|/\|\mathbf{x}_t - \mathbf{x}^\star\| \to 0$ as $t \to \infty$. However, the analyses in [31, 50–52] highly rely on specific properties of quasi-Newton updates in the Broyden class, and do not apply to general Hessian approximations (e.g., finite difference approximation).

1.1 Stochastic Newton methods

A parallel line of research explores stochastic Newton methods, where a stochastic approximation $\hat{\mathbf{H}}_t$ is used in place of $\mathbf{H}_t$ in (2). The stochasticity of $\hat{\mathbf{H}}_t$ may come from evaluating the Hessian on a random subset of data points (i.e., subsampling), or from projecting the Hessian onto a random subspace to achieve the dimension reduction (i.e., sketching). To unify different approaches, we consider in this paper a general Hessian approximation given by a *stochastic oracle*. In particular, we express the approximation $\hat{\mathbf{H}}(\mathbf{x})$ at any point $\mathbf{x}$ by

$$\hat{\mathbf{H}}(\mathbf{x}) = \mathbf{H}(\mathbf{x}) + \mathbf{E}(\mathbf{x}), \tag{4}$$

where $\mathbf{E}(\mathbf{x}) \in \mathbb{S}^{d\times d}$ is a symmetric random noise matrix following a certain distribution (conditional on $\mathbf{x}$) with mean zero. At iterate $\mathbf{x}_t$, we query the oracle to obtain an approximation $\hat{\mathbf{H}}_t = \hat{\mathbf{H}}(\mathbf{x}_t)$, which (implicitly) comes from generating a realization of the random matrix $\mathbf{E}_t = \mathbf{E}(\mathbf{x}_t)$, and then adding $\mathbf{E}_t$ to the true Hessian $\mathbf{H}_t$. We do not assume $\mathbf{E}_t$ and $\mathbf{H}_t$ are accessible to us.

As mentioned, the popular specializations of stochastic oracle include subsampling and sketching. For Hessian subsampling, a finite-sum objective is considered

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}). \tag{5}$$

In the t-th iteration, a subset $\xi_t \subseteq \{1, 2, \ldots, n\}$ is sampled uniformly at random, and the subsampled Hessian is defined as

$$\hat{\mathbf{H}}_t = \frac{1}{|\xi_t|}\sum_{i\in\xi_t} \nabla^2 f_i(\mathbf{x}_t). \tag{6}$$

We note that the components f_i in (5) may not be convex even if the function f is strongly convex.[1] This is the so-called *sum-of-non-convex* setting [2, 25, 26, 55, e.g., see]. Our oracle model (4) allows for this, since $\hat{\mathbf{H}}_t$ is not required to be positive semidefinite, while the existing Subsampled Newton methods generally do not allow it. See Sect. 1.3 and Example 1 for further discussion.

For Hessian sketching, one first forms the square-root Hessian matrix $\mathbf{M}_t \in \mathbb{R}^{n\times d}$ satisfying $\mathbf{H}_t = \mathbf{M}_t^\top \mathbf{M}_t$, where n is the number of data points. Then, one generates a random sketch matrix $\mathbf{S}_t \in \mathbb{R}^{s\times n}$ with the sketch size s and the property $\mathbb{E}[\mathbf{S}_t^\top \mathbf{S}_t] = \mathbf{I}$, and the sketched Hessian is defined as

$$\hat{\mathbf{H}}_t = \mathbf{M}_t^\top \mathbf{S}_t^\top \mathbf{S}_t \mathbf{M}_t. \tag{7}$$

In some cases, $\mathbf{M}_t$ can be easily obtained. One particular example is a generalized linear model, where the objective has the form

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{a}_i^\top \mathbf{x})$$

with $\{\mathbf{a}_i\}_{i=1}^n \in \mathbb{R}^d$ being n data points. In this case, $\mathbf{M}_t = \frac{1}{\sqrt{n}} \cdot \mathrm{diag}(f_1''(\mathbf{a}_1^\top \mathbf{x}_t)^{1/2}, \ldots, f_n''(\mathbf{a}_n^\top \mathbf{x}_t)^{1/2})\mathbf{A}$ where $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)^\top \in \mathbb{R}^{n\times d}$ is the data matrix.

Another family of stochastic Newton methods is based on the so-called Sketch-and-Project framework (e.g., [28]), where a low-rank Hessian estimate is used to construct a Newton-like step (with the Moore-Penrose pseudoinverse instead of the matrix inverse). For example, in one approach [27], the Newton-like step in the t-th iteration is obtained by generating a sketching matrix $\mathbf{S}_t \in \mathbb{R}^{s\times d}$ to construct a rank-s approximate of the inverse Hessian, resulting in the update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mu_t \mathbf{S}_t^\top (\mathbf{S}_t \mathbf{H}_t \mathbf{S}_t^\top)^\dagger \mathbf{S}_t \nabla f(\mathbf{x}_t).$$

[1] A concrete example is finding the leading eigenvector of a covariance matrix $\mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top$. Here, the objective can be structured as $f_{\boldsymbol{b},\nu}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}^\top(\nu\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\top)\mathbf{x} - \boldsymbol{b}^\top\mathbf{x})$, where $\nu > \|\mathbf{A}\|$ and $\boldsymbol{b}$ are given. See [25] for details.

While linear convergence rates have been derived for these methods (e.g., [16]), they do not fit into our stochastic oracle framework due to an intrinsic bias arising in the Hessian estimates (see Sect. 1.3 for further discussion).

Numerous stochastic Newton methods using (6) or (7) with various types of convergence guarantees have been proposed [1, 6, 10, 11, 15, 17, 18, 22, 24, 32, 34, 48, 53]. We review these related references later and point to [4] for a recent brief survey. However, the existing approaches to stochastic Newton methods share common limitations, which we now discuss.

The majority of literature studies the convergence of stochastic Newton by establishing a *high probability error recursion*. For example, [1, 22, 48, 53] all showed, roughly speaking, that stochastic Newton methods exploit a local linear-quadratic recursion:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq c_1\|\mathbf{x}_t - \mathbf{x}^\star\| + c_2\|\mathbf{x}_t - \mathbf{x}^\star\|^2 \qquad \text{with probability } 1-\delta, \tag{8}$$

for some constants $c_1, c_2 > 0$ and $\delta \in (0, 1)$. These constants depend on the sample/sketch size at each step. Based on (8), one often applies the result for $t = 0, 1, \ldots, T$ recursively, uses the union bound, and establishes local convergence with probability $1 - T\delta$. This approach has the following key drawbacks.

(a) The presence of the linear term with coefficient $c_1 > 0$ in the recursion means that, once $\mathbf{x}_t$ is sufficiently close to $\mathbf{x}^\star$, the algorithm can only achieve the linear convergence, as opposed to the quadratic or superlinear convergence achieved by deterministic methods. Prior works (e.g., [6, 53]) discussed how to achieve local superlinear convergence by diminishing the coefficient $c_1 = c_{1,t}$ gradually as t increases. However, since c_1 is proportional to the magnitude of stochastic noise in the Hessian estimates, diminishing it requires increasing the sample size for estimating the Hessian, which results in a blow-up of the per-iteration computational cost. To be specific, c_1 is proportional to the reciprocal of the square root of the sample size; thus this blow-up in terms of the sample size can be as fast as exponential if we wish to attain the quadratic convergence rate, and linear if we wish to attain the superlinear convergence rate established in this paper later.

(b) The presence of the failure probability δ in (8) means that after T iterations, a convergence guarantee may only hold with probability $1 - T\delta$. Thus, the failure probability explodes when $T \to \infty$. To resolve this issue one can gradually diminish δ, e.g., $\delta = \delta_t = O(1/t^2)$ such that $\sum_t \delta_t < \infty$. However, once again, such adjustment on δ leads to an increasing sample/sketch size as the algorithm proceeds, although not as drastically as in (a) (it suffices to increase the sample size logarithmically). Thus, in our stochastic oracle model, where the noise level (and hence the sample size) remains constant, it is problematic to show any convergence with high probability from (8) as $T \to \infty$.

We note that some prior works [6, 42] showed the convergence in expectation guarantees. Although the explosion of the failure probability in (b) is suppressed by the direct characterization of the expectation, the drawback (a) still remains. In addition, the convergence in expectation results require stronger assumptions. For example, [6, (2.17)] and [42, Theorem 2] showed similar recursions to (8) for $\mathbb{E}[\|\mathbf{x}_t -$

$\mathbf{x}^\star\|]$. However, these works assumed a *bounded moments of iterates* condition, i.e., $\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^\star\|^2] \leq \gamma(\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^\star\|])^2$ for some constant $\gamma > 0$, and assumed each component f_i in (5) to be strongly convex. Both conditions are not needed in high probability analyses. Note that the majority of high probability analyses assume that each component f_i is convex though not necessarily strongly convex [1, 53], while we do not impose any conditions on f_i, which significantly weakens the conditions in the existing literature. Furthermore, the stepsize in [6, 42] was prespecified without any adaptivity. [3] introduced a non-monotone line search to address the adaptivity issue, while extra conditions such as the compactness of $\{\mathbf{x}_t\}_t$ were imposed for their expectation analysis.

1.2 Main results: stochastic Newton with Hessian averaging

Concerned by the above limitations of stochastic Newton methods, we ask the following question:

> Can we design a stochastic Newton method that exploits global linear and local superlinear convergence in high probability, even for an infinite iteration sequence, and without increasing the computational cost in each iteration?

We provide an affirmative answer to this question by studying a class of stochastic Newton methods with Hessian averaging. A simple intuition is that the approximation error $\mathbf{E}_t$ can be de-noised by aggregating all the errors $\{\mathbf{E}_i\}_{i=0}^t$, inspired by the central limit theorem for martingale differences. Thus, if we could reuse the past samples and replace $\mathbf{E}_t$ by $\frac{1}{t+1}\sum_{i=0}^t \mathbf{E}_i$, then the matrix $\mathbf{H}_t + \frac{1}{t+1}\sum_{i=0}^t \mathbf{E}_i$ would be more precise than $\hat{\mathbf{H}}_t = \mathbf{H}_t + \mathbf{E}_t$. However, since $\{\mathbf{E}_i\}_{i=0}^t$ are unknown and only $\{\hat{\mathbf{H}}_i\}_{i=0}^t$ are known to us, in order to de-noise $\mathbf{E}_t$, we can only aggregate all the Hessian estimates $\{\hat{\mathbf{H}}_i\}_{i=0}^t$ and derive $\frac{1}{t+1}\sum_{i=0}^t \hat{\mathbf{H}}_i = \frac{1}{t+1}\sum_{i=0}^t \mathbf{H}_i + \frac{1}{t+1}\sum_{i=0}^t \mathbf{E}_i$. Compared to $\hat{\mathbf{H}}_t$, such a matrix does not preserve the information of the true Hessian $\mathbf{H}_t$. Thus, we observe a trade-off between de-noising the error $\mathbf{E}_t$ and preserving the Hessian information $\mathbf{H}_t$: on one hand, we want to assign equal weights to all the past errors to achieve the fastest concentration for the error average (see Remark 2); on the other hand, we want to assign all weights to $\mathbf{H}_t$ to fully preserve the most recent Hessian information. In this paper, we investigate this trade-off, and show how the Hessian averaging resolves the drawbacks of existing stochastic Newton methods.

1.2.1 The proposed method

We consider an online averaging scheme (we let $\tilde{\mathbf{H}}_{-1} = \mathbf{0}$):

$$\tilde{\mathbf{H}}_t = \frac{w_{t-1}}{w_t}\tilde{\mathbf{H}}_{t-1} + \left(1 - \frac{w_{t-1}}{w_t}\right)\hat{\mathbf{H}}_t, \quad t = 0, 1, 2, \ldots, \tag{9}$$

where $\{w_t\}_{t=-1}^\infty$ is a prespecified increasing non-negative weight sequence starting with $w_{-1} = 0$. By *online* we mean that we only keep an average Hessian $\tilde{\mathbf{H}}_t$ in each iteration, and update it as (9) when we obtain a new Hessian estimate. This scheme

Algorithm 1 Stochastic Newton Method with Hessian Averaging

1: **Input**: iterate $\mathbf{x}_0$, weights $\{w_t\}_{t=-1}^{\infty}$, scalars $\beta \in (0, 1/2)$, $\rho \in (0, 1)$, and $\tilde{\mathbf{H}}_{-1} = \mathbf{0}$;
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Obtain a stochastic Hessian approximation $\hat{\mathbf{H}}_t = \mathbf{H}_t + \mathbf{E}_t$;
4: Compute the Hessian average via (9): $\tilde{\mathbf{H}}_t = \frac{w_{t-1}}{w_t}\tilde{\mathbf{H}}_{t-1} + \left(1 - \frac{w_{t-1}}{w_t}\right)\hat{\mathbf{H}}_t$;
5: Compute the Newton direction $\mathbf{p}_t$ by solving: $\tilde{\mathbf{H}}_t \mathbf{p}_t = -\nabla f_t$;
6: **if** $\tilde{\mathbf{H}}_t \mathbf{p}_t = -\nabla f_t$ is not solvable **or** $\nabla f_t^\top \mathbf{p}_t \geq 0$, **then** skip iteration with $\mathbf{x}_{t+1} = \mathbf{x}_t$;
7: Compute a stepsize $\mu_t = \rho^{j_t}$, where j_t is the smallest nonnegative integer such that

$$\text{Armijo condition :} \qquad f(\mathbf{x}_t + \mu_t \mathbf{p}_t) \leq f(\mathbf{x}_t) + \mu_t \beta \nabla f_t^\top \mathbf{p}_t$$

8: Update $\mathbf{x}_{t+1} = \mathbf{x}_t + \mu_t \mathbf{p}_t$;
9: **end for**

is in contrast to keeping all the past Hessian estimates. By the scheme (9), we note that, at iteration t, we re-scale the weights of $\{\hat{\mathbf{H}}_i\}_{i=0}^{t-1}$ equally by a factor of w_{t-1}/w_t, instead of completely redefining all the weights. We note that such an online averaging scheme is memory and computation efficient: compared to stochastic Newton methods without averaging, we only require an extra $O(d^2)$ space to save $\tilde{\mathbf{H}}_t$ and $O(d^2)$ flops to update it, which is negligible in the setting of stochastic Newton methods where the Hessian vector product requires $O(d^2)$ flops and solving the exact Newton system requires $O(d^3)$ flops. In (9), we use the ratio factors w_{t-1}/w_t instead of direct weights merely to simplify our later presentation. Furthermore, (9) can be reformulated as a general weighted average as follows:

$$\tilde{\mathbf{H}}_t = \sum_{i=0}^{t} z_{i,t} \hat{\mathbf{H}}_i, \qquad \text{with} \quad z_{i,t} := (w_i - w_{i-1})/w_t. \tag{10}$$

In particular, by setting $w_t = t + 1$, we obtain $z_{i,t} = 1/(t + 1)$ and further have $\tilde{\mathbf{H}}_t = \frac{1}{t+1}\sum_{i=0}^{t} \hat{\mathbf{H}}_i$. Thus, we recover simple uniform averaging. If we let the sequence w_t grow faster-than-linearly, this results in a weighted average that is skewed towards the more recent Hessian estimates.

Our proposed method replaces $\mathbf{H}_t$ by $\tilde{\mathbf{H}}_t$ when computing the Newton direction (2). The detailed procedure is displayed in Algorithm 1. We make two comments about the algorithm.

First, Algorithm 1 supposes that, unlike the Hessian, the function values and gradients are known deterministically. As a result, the method generates a monotonically decreasing sequence of $f(\mathbf{x}_t)$. If, on the other hand, $f(\mathbf{x}_t)$ and $\nabla f(\mathbf{x}_t)$ were known with random noise, we would have to relax the Armijo condition by adding extra error terms (hence, $f(\mathbf{x}_t)$ would be potentially non-monotonic), and analyze the resulting method under a random model framework such as in [5, 12]. We leave these non-trivial extensions to future work.

Second, for early iterates, the average Hessian $\tilde{\mathbf{H}}_t$ may not be a good approximate of $\mathbf{H}_t$. For example, $\tilde{\mathbf{H}}_t$ may be indefinite (or even singular) so that $\mathbf{p}_t$ is not a descent direction. In that case, we skip the line search step and let $\mathbf{x}_{t+1} = \mathbf{x}_t$ (see Line

6 of Algorithm 1). Further, even if $\tilde{\mathbf{H}}_t$ is positive definite, it may not be properly bounded, and thus leads to an extremely small stepsize μ_t. However, as analyzed later in Lemmas 2 and 7, the errors $\{\mathbf{E}_i\}_{i=0}^t$ are sufficiently concentrated for large t with high probability, so that $\tilde{\mathbf{H}}_t$ is properly bounded from above and below. Thus, Algorithm 1 is well defined and can be performed for any iteration t, while it behaves like the classical Newton method only after a few iterations (the threshold is provided in Lemmas 2 and 7).

1.2.2 Main results

We conduct convergence analysis for Algorithm 1 with different weight sequences $\{w_t\}_{t=-1}^{\infty}$. Throughout the analysis, we only assume the oracle noise $\mathbf{E}(\mathbf{x})$ has a sub-exponential tail, which in particular includes Hessian subsampling and Hessian sketching as special cases. Our convergence guarantees rely on the quality of the average Hessian approximation $\tilde{\mathbf{H}}_t$; thus, we do not require $\hat{\mathbf{H}}_t$ to be a good approximation of $\mathbf{H}_t$. In other words, our scheme is applicable even if we generate a *single* sample for forming the subsampled Hessian estimator, and applicable even if some components f_i are non-convex. This is because the noise $\mathbf{E}_t$ can always be reduced by averaging (ensured by the central limit theorem) even if it has a large variance.

We show that, with high probability, Algorithm 1 has four convergence phases with three transition points:

(a) Global phase: $\mathbf{x}_t$ converges from any initial iterate $\mathbf{x}_0$ to a local neighborhood of $\mathbf{x}^\star$, in which the unit stepsize starts being accepted.
(b) Steady phase: $\mathbf{x}_t$ stays in the neighborhood.
(c) Slow superlinear phase: $\mathbf{x}_t$ starts converging superlinearly with a rate gradually increasing.
(d) Fast superlinear phase: the superlinear acceleration reaches its full potential and is maintained for all the following iterations.

We mention that the superlinear rate is measured with respect to the error $\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} := \sqrt{(\mathbf{x}_t - \mathbf{x}^\star)^\top \mathbf{H}^\star (\mathbf{x}_t - \mathbf{x}_0)}$ where $\mathbf{H}^\star = \mathbf{H}(\mathbf{x}^\star)$. Before introducing the main results in Theorems 1 and 2, we summarize them in Table 1, showing the transition points for two weight sequences. The transitions for general weights are provided in Sect. 4. We use Υ to denote the *noise level* of stochastic Hessian oracle (see Assumption 1 for a formal definition; typically $\Upsilon = O(\kappa)$). We also use $O(\cdot)$ to suppress the logarithmic factors and the dependence on other constants except Υ and κ. We emphasize that all results of the paper require that the (weighted) average of the errors $\{\mathbf{E}_i\}_{i=0}^t$ is sufficiently concentrated; thus, they hold *with high probability*.

The first main result studies the uniform averaging scheme, which is informally stated below. We refer to Theorem 4 for a formal statement.

Theorem 1 (Uniform averaging; informal) *Consider Algorithm 1 with $w_t = t + 1$. With high probability, the algorithm satisfies:*

1. *After $\mathcal{T} = O(\kappa^2 + \Upsilon^2)$ iterations, $\mathbf{x}_t$ converges to a local neighborhood of $\mathbf{x}^\star$.*
2. *After $O(\mathcal{T}\kappa)$ iterations, $\mathbf{x}_t$ converges superlinearly to $\mathbf{x}^\star$.*

Table 1 Three transitions of two averaging schemes: uniform averaging ($w_t = t+1$, see Theorem 1) and our proposed weighted averaging scheme ($w_t = (t+1)^{\log(t+1)}$, see Theorem 2)

Weights w_t	First transition	Second transition	Third transition
$t+1$ (Thm. 1)	$O(\kappa^2+\Upsilon^2)$	$O(\kappa\cdot\{\kappa^2+\Upsilon^2\})$	$O(\kappa^2/\Upsilon^2\cdot\{\kappa^2+\Upsilon^2\}^2)$
$(t+1)^{\log(t+1)}$ (Thm. 2)	$O(\kappa^2+\Upsilon^2)$	$O(\kappa^2+\Upsilon^2)$	$O(\kappa^2+\Upsilon^2)$

For each weight sequence, the three transitions occur with high probability

3. *After $O(\mathcal{T}^2\kappa^2/\Upsilon^2)$ iterations, the superlinear rate reaches $(\Upsilon\sqrt{\log(t)/t}\,)^t$, and this rate is maintained as $t\to\infty$.*

First, we observe that our convergence guarantee holds with high probability for the entire infinite iteration sequence, which addresses the issue of the blow-up of failure probability associated with the existing stochastic Newton methods (see part (b) in Sect. 1.1).

Second, the parameter Υ characterizes the noise level of the stochastic Hessian oracle. When the Hessian is generated by subsampling or sketching, Υ depends on the adopted sample/sketch sizes. As illustrated in Examples 1 and 2, $\Upsilon = O(\kappa)$ for popular Hessian estimators when sample/sketch sizes are independent of κ. In this case, $\mathcal{T} = O(\kappa^2)$. We require $\mathcal{T} \geq O(\Upsilon^2)$ only to ensure that $\{\tilde{\mathbf{H}}_t\}_{t\geq\mathcal{T}}$ are positive definite with condition numbers scaling as κ, so that the Newton step based on $\tilde{\mathbf{H}}_t$ leads to a usual decrease of the objective. On the other hand, popular stochastic Newton methods often generate a larger sample size (which depends on κ) to enforce $\|\mathbf{E}_t\| \leq \epsilon\lambda_{\min}$ for any $t \geq 0$ with an $\epsilon \in (0,1)$ (e.g., see Lemma 2 and Equation 3.10 in [48, 53] respectively). In that case, $\{\hat{\mathbf{H}}_t\}_{t\geq 0}$ are positive definite and so are $\{\tilde{\mathbf{H}}_t\}_{t\geq 0}$. Importantly, our method does not require such well-conditioned Hessian oracles.

Third, we notice that the uniform averaging approach has three transitions outlined in Theorem 1. For the iterations before $\mathcal{T}$, the error in function value decreases linearly, implying that $\mathbf{x}_t$ converges R-linearly. The *first transition* occurs after $\mathcal{T}$ iterations when $\mathbf{x}_t$ reaches the local neighborhood, where second-order information starts being useful. $\mathcal{T}$ is also where the exact Newton methods would reach quadratic convergence. However, the averaged Hessian estimates still carry inaccurate Hessian information from the global phase, which is only gradually forgotten. From $\mathcal{T}$ to $O(\mathcal{T}\kappa)$ iterations, the algorithm gradually forgets the Hessian estimates in the global phase, while $\mathbf{x}_t$ still converges R-linearly. The *second transition* occurs after $O(\mathcal{T}\kappa)$ iterations, when $\mathbf{x}_t$ starts converging superlinearly (with a slow rate). The *third transition* occurs after $O(\mathcal{T}^2\kappa^2/\Upsilon^2)$, when the superlinear rate is accelerated to $(\Upsilon\sqrt{\log(t)/t}\,)^t$ and stabilized. This rate comes from the central limit theorem of averaging out the oracle noise; thus, this rate cannot be further improved.

Note that if we were to reset the averaged Hessian estimate $\tilde{\mathbf{H}}_t$ after the first transition (so that the Hessian average does not include information from the global phase), then the algorithm would immediately reach the superlinear rate $(\Upsilon\sqrt{\log(t)/t}\,)^t$ after $\mathcal{T}$ iterations (i.e., all transitions would occur at once). However, the algorithm does not know a priori when this transition will occur. As a result, the uniform Hessian averaging

incurs a potentially significant delay of up to $O(\mathcal{T}^2\kappa^2/\Upsilon^2)$ iterations before reaching the desired superlinear rate.

In our second main result, we address the delayed transition to superlinear convergence that occurs in the uniform averaging. Specifically, we ask:

> Does there exist a universal weighted averaging scheme that achieves superlinear convergence without a delayed transition, and without any prior knowledge about the objective?

Remarkably, the answer to this question is affirmative. The weighted/non-uniform averaging scheme we present below puts more weight on recent Hessian estimates, so that the second-order information from the global phase is forgotten more quickly, and the transition to a fast superlinear rate occurs after only $O(\mathcal{T})$ iterations. Thus, the superlinear convergence occurs without any delay (up to constant factors). Such a scheme will necessarily have a slightly weaker superlinear rate than the uniform averaging as $t \to \infty$, but we show that this difference is merely an additional $O(\sqrt{\log t})$ factor (see Theorem 5 and Example 5 for a formal statement).

Theorem 2 (Weighted averaging; informal) *Consider Algorithm 1 with* $w_t = (t+1)^{\log(t+1)}$. *With high probability, after* $O(\mathcal{T}) = O(\kappa^2 + \Upsilon^2)$ *iterations,* $\mathbf{x}_t$ *achieves a superlinear convergence rate* $(\Upsilon \log(t)/\sqrt{t})^t$, *which is maintained as* $t \to \infty$.

We note that, given some knowledge about the global/local transition points (e.g., if the algorithm knows κ, or if some convergence criterion is used for estimating the transition point), it is possible to switch from the more conservative weighted averaging to the asymptotically more effective uniform averaging within one run of the algorithm. However, since knowing transition points is difficult and rare in practice, we leave such considerations to future work, and only focus on problem-independent averaging schemes.

It is also worth mentioning that this paper only considers a basic stochastic Newton scheme based on (2), where we suppose exact function and gradient information and solutions to the Newton systems are known exactly. Some literature allows one to access inexact function values and/or gradients, and/or apply Newton-CG or MINRES to solve the linear systems inexactly [23, 37, 53, 63]. Applying our sample aggregation technique under these setups is promising, but we defer it to future work. The basic scheme purely reflects the benefits of Hessian averaging, which is the main interest of this work. Additionally, some literature deals with non-convex objectives via stochastic trust region methods [5, 12] or stochastic Levenberg-Marquardt methods [39]. The averaging scheme may not directly apply for these methods due to potential bias brought by Hessian modifications for addressing non-convexity, while our sample aggregation idea is still inspiring. We leave the generalization to non-convex objectives to future work as well. Some literature addressed the superlinearity of stochastic Newton methods under distributed or federated learning settings [30, 49, 54]. These works are not fully compatible with our Hessian oracle framework, since they exploit some distributed nature of problem to produce Hessian estimates with noise diminishing to zero (as opposed to the bounded noise in this paper).

1.3 Literature review

Stochastic Newton methods have recently received much attention. The popular Hessian approximation methods include subsampling and sketching.

For subsampled Newton methods, aside from extensive empirical studies on different problems [33, 40, 41, 62], the pioneering work in [10] established the very first asymptotic global convergence by showing that $\|\nabla f_t\| \to 0$ as $t \to \infty$, while the quantitative rate is unknown. Furthermore, [11, 24] studied Newton-type algorithms with subsampled gradients and/or subsampled Hessians, and established global Q-linear convergence in the error of function value $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$. However, the above analyses neglected the underlying probabilistic nature of the subsampled Hessian $\hat{\mathbf{H}}_t$, and required $\hat{\mathbf{H}}_t$ to be lower bounded away from zero *deterministically*. Such a condition holds only if each f_i in (5) is strongly convex, which is restrictive in general. Erdogdu and Montanari [22] relaxed such a condition by developing a novel algorithm, where the subsampled Hessian is adjusted by a truncated eigenvalue decomposition. With the exact gradient information and properly prespecified stepsizes, the authors showed a linear-quadratic error recursion for $\|\mathbf{x}_t - \mathbf{x}^\star\|$ in high probability. Arguably, the convergence of standard subsampled Newton methods is originally analyzed in [53] and [6] from different perspectives. In particular, for both sampling and not sampling the gradient, [53] showed a global Q-linear convergence for $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ and a local linear-quadratic convergence for $\|\mathbf{x}_t - \mathbf{x}^\star\|$ in high probability. Under some additional conditions, [6] derived a global R-linear convergence for the expected function value $\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)]$ and a (similar) local linear-quadratic convergence for the expected iterate error $\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^\star\|]$. For both works, the authors also discussed how to gradually increase the sample size for Hessian approximation to achieve a local Q-superlinear convergence with high probability and in expectation, respectively. Building on the two studies, various modifications of subsampled Newton methods have been reported with similar convergence guarantees. We refer to [3, 36, 61, 64] and references therein. We note that [32] designed a scheme that allows for a single sample in each iteration of subsampled Newton. That work established a local linear convergence in expectation, while we obtain a superlinear convergence in high probability.

As a parallel approach to subsampling, Newton sketch has also been broadly investigated. Pilanci and Wainwright [48] proposed a generic Newton sketch method that approximates the Hessian via a Johnson–Lindenstrauss (JL) transform (e.g., the Hadamard transform), and the gradient is exact. Furthermore, [1, 15, 17, 18] proposed different Newton sketch methods with debiased or unbiased Hessian inverse approximations. Dereziński et al. [19] relied on a novel sketching technique called Leverage Score Sparsified (LESS) embeddings [20] to construct a sparse sketch matrix, and studied the trade-off between the computational cost of $\hat{\mathbf{H}}_t$ and the convergence rate of the algorithm. Similar to subsampled Newton methods, the aforementioned literature established a local linear-quadratic (or linear) recursion for $\|\mathbf{x}_t - \mathbf{x}^\star\|$ in high probability. A recent work [34] adaptively increased the sketch size to let the linear coefficient be proportional to the iterate error, which leads to a quadratic convergence. However, the per-iteration computational cost is larger than typical methods.

See [4] for a review of subsampled and sketched Newton, and their connections to, and empirical comparisons with, first-order methods.

Sketching has also been used to construct *low-rank* Hessian approximations through the Sketch-and-Project framework, originally developed by [28] for solving linear systems, and extended to general convex/nonconvex optimization by [21, 27, 38, 44]. The convergence properties of this family of methods have been thoroughly studied: they achieve linear convergence in expectation, with the rate controlled by a so-called *stochastic condition number*, which is defined as the smallest eigenvalue of the expectation of the low-rank projection matrix defined by the sketch [28]. While the per-iteration cost of stochastic Newton methods based on Sketch-and-Project is generally lower than that of the aforementioned Newton sketch methods, their convergence rates are more sensitive to the spectral properties of the Hessian. The precise characterizations of the convergence rates are given in [16, 18, 43]. Moreover, the Sketch-and-Project estimates are generally biased, so they are not appropriate for averaging.

In summary, none of the aforementioned existing works achieve superlinear convergence with a fixed per-iteration computational cost. Additionally, high probability convergence guarantees generally fail as $t \to \infty$, with potent exceptions of certain stochastic trust-region methods [12] that enjoy almost sure convergence. However, the per-iteration computation of the exceptions is not fixed and the local rate is unknown. Further, for finite-sum objectives, the existing literature on stochastic Newton methods assumes each f_i to be strongly convex [6] or convex [53]. However, f_i needs not be convex even if f is strongly convex. See [2, 25, 26, 55] and references therein for first-order algorithms designed under such a setting. We address the above limitations of stochastic Newton methods by reusing all the past samples to average Hessian estimates. Our scheme is especially preferable when we have a limited budget for per-iteration computation (e.g., when we use very few samples in subsampled Newton, resulting in an ill-conditioned Hessian estimate). Our established non-asymptotic superlinear rates are stronger than the existing results, and our numerical experiments demonstrate the superiority of Hessian averaging.

Notation: Throughout the paper, we use $\mathbf{I}$ to denote the identity matrix, and $\mathbf{0}$ to denote the zero vector or matrix. Their dimensions are clear from the context. We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and spectral norm for matrices. For a positive semidefinite matrix $\mathbf{A}$, we let $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. For two scalars $a, b, a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two matrices $\mathbf{A}, \mathbf{B}$, $\mathbf{A} \prec (\preceq)\mathbf{B}$ if $\mathbf{B} - \mathbf{A}$ is a positive (semi)definite matrix. Recall that we reserve the notation $\lambda_{\min}, \lambda_{\max}$ to denote the lower and upper bounds of the true Hessian, and $\kappa = \lambda_{\max}/\lambda_{\min}$ is the condition number.

Structure of the paper: In Sect. 2 we present the preliminaries on matrix concentration that are needed for our results. Then, we establish convergence results for the uniform averaging scheme in Sect. 3. Section 4 establishes convergence for general weight sequences. Numerical experiments and conclusions are provided in Sects. 5 and 6, respectively.

2 Preliminaries on matrix concentration

In this section, we present the key results on the concentration of sums of random matrices, which we then use to bound the noise of the averaged Hessian estimates. The Hessian estimates constructed by our algorithm are not independent, and hence standard matrix concentration results do not apply. However, they do satisfy a martingale difference condition, which we will exploit to derive useful concentration results.

Given a sequence of stochastic iterates $\{\mathbf{x}_t\}_{t=0}^{\infty}$, we let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$ be a filtration where $\mathcal{F}_t = \sigma(\mathbf{x}_{0:t})$, $\forall t \geq 0$, is the σ-algebra generated by the randomness from $\mathbf{x}_0$ to $\mathbf{x}_t$. With such a filtration, we denote the conditional expectation by $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and the conditional probability by $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$. We suppose $\mathbf{x}_0$ is deterministic, so that $\mathcal{F}_0$ is a trivial σ-algebra.

For a given weight sequence $\{w_t\}_{t=-1}^{\infty}$ with $w_{-1} = 0$, the scheme (9) leads to

$$\tilde{\mathbf{H}}_t \overset{(10)}{=} \sum_{i=0}^{t} z_{i,t}\hat{\mathbf{H}}_i \overset{(4)}{=} \underbrace{\sum_{i=0}^{t} z_{i,t}\mathbf{H}_i}_{\text{Hessian averaging}} + \bar{\mathbf{E}}_t \qquad \text{for} \quad \bar{\mathbf{E}}_t := \underbrace{\sum_{i=0}^{t} z_{i,t}\mathbf{E}_i}_{\text{noise averaging}} \tag{11}$$

where $z_{i,t} = (w_i - w_{i-1})/w_t$. Note that $\sum_{i=0}^{t} z_{i,t} = 1$ and $z_{i,t} \propto z_i := w_i - w_{i-1}$, i.e., $z_{i,t}$ is proportional to an un-normalized weight z_i. We see from (11) that $\tilde{\mathbf{H}}_t$ consists of the Hessian averaging and noise averaging with the same weights. In principle, the Hessian averaging $\sum_{i=0}^{t} z_{i,t}\mathbf{H}_i \to \mathbf{H}^\star$ as $\mathbf{x}_t \to \mathbf{x}^\star$, while the noise averaging $\sum_{i=0}^{t} z_{i,t}\mathbf{E}_i \to \mathbf{0}$ due to the central limit theorem. We will show that the Hessian averaging (eventually) converges faster than the noise averaging.

To study the concentration of noise averaging $\bar{\mathbf{E}}_t$, we use the fact that $\{\mathbf{E}_t\}_{t=0}^{\infty}$ is a martingale difference sequence, and rely on concentration inequalities for matrix martingales. These concentration inequalities require a sub-exponential tail condition on the noise. We say that a random variable X is K-sub-exponential if $\mathbb{E}[|X|^p] \leq p! \cdot K^p/2$ for all $p = 2, 3, \dots$, which is consistent (up to constants) with all standard notions of sub-exponentiality (see Sect. 2.7 in [59]).

Assumption 1 (*Sub-exponential noise*) We assume that $\mathbf{E}(\mathbf{x})$ is mean zero and $\|\mathbf{E}(\mathbf{x})\|$ is Υ_E-sub-exponential for all $\mathbf{x}$. Also, we define $\Upsilon := \Upsilon_E/\lambda_{\min}$ to be the scale-invariant noise level.

Remark 1 The sub-exponentiality of $\|\mathbf{E}(\mathbf{x})\|$ implies that $\mathbf{E}(\mathbf{x})$ has sub-exponential matrix moments: $\mathbb{E}[\mathbf{E}(\mathbf{x})^p] \preceq p! \cdot \Upsilon_E^p/2 \cdot \mathbf{I}$ for $p = 2, 3, \dots$. In fact, our analysis immediately applies under this slightly weaker condition. We impose the moment condition on $\|\mathbf{E}(\mathbf{x})\|$ purely because it is easier to check in practice. Also, we note that sometimes the noise $\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}$ is more natural to study, e.g., for sketching-based oracles where we additionally have $\hat{\mathbf{H}}(\mathbf{x}) \succeq \mathbf{0}$. Thus, we can alternatively impose $\tilde{\Upsilon}$-sub-exponentiality on $\|\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}\|$. Our analysis can also be adapted to this alternate condition, and leads to tighter convergence rates (in terms of the dependence on κ) for particular sketching-based oracles. However, this adaptation loses certain generality, and thus we prefer to impose conditions directly on the oracle noise $\mathbf{E}(\mathbf{x})$.

Assumption 1 is weaker than assuming $\|\mathbf{E}(\mathbf{x})\|$ to be uniformly bounded by Υ_E, and it is satisfied by all of the popular Hessian subsampling and sketching methods. For example, when using Gaussian sketching, the noise is not bounded but sub-exponential. Further, the sub-exponential constant Υ_E (and hence Υ) reflects how the stochastic noise depends on the sample/sketch size, as illustrated in the examples below (see Appendix A for rigorous proofs).

Example 1 Consider subsampled Newton as in (6) with sample size $s = |\xi_t|$, and suppose $\|\nabla^2 f_i(\mathbf{x})\| \leq \lambda_{\max} R$ for some $R > 0$ and for all i. Then, we have $\Upsilon = O(\kappa R\sqrt{\log(d)/s})$. If we additionally assume that all $f_i(\mathbf{x})$ are convex, then Υ is improved to $\Upsilon = O(\sqrt{\kappa R \log(d)/s} + \kappa R \log(d)/s)$.

Example 2 Consider Newton sketch as in (7) with $\mathbf{S} \in \mathbb{R}^{s\times n}$ consisting of i.i.d. $\mathcal{N}(0, 1/s)$ entries. Then, we have $\Upsilon = O(\kappa(\sqrt{d/s} + d/s))$.

From the above two examples, we observe that Υ scales as $O(\kappa)$ when holding everything else fixed. Also, Example 1 illustrates that our Hessian oracle model applies to subsampled Newton even when some components $f_i(\mathbf{x})$ are non-convex (while $f(\mathbf{x})$ is still convex), although this adversely affects the sub-exponential constant. For Gaussian sketch in Example 2, we can show that $\widetilde{\Upsilon} = O(\sqrt{d/s} + d/s)$ (where $\widetilde{\Upsilon}$ was defined in Remark 1). Thus, the dependence on κ can be avoided for the sub-exponential constant of noise $\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}$. Analogous noise bounds can be proved for other sketching matrices $\mathbf{S}$, including sparse sketches and randomized orthogonal transforms.

We now show concentration inequalities for $\bar{\mathbf{E}}_t$ in (11) under Assumption 1. We state the following preliminary lemma, which is a variant of Freedman's inequality for matrix martingales.

Lemma 1 (Adapted from Theorem 2.3 in [57]) *Let $t \geq 0$ be a fixed integer. Consider a d-dimensional martingale difference $\{\mathbf{E}_i\}_{i=0}^t$ (i.e., $\mathbb{E}_i[\mathbf{E}_i] = \mathbf{0}$). Suppose there exists a function $g_t : \Theta_t \to [0, \infty]$ with $\Theta_t \subseteq (0, \infty)$, and a sequence of matrices $\{\mathbf{U}_i\}_{i=0}^t$, such that for any $i = 0, 1, \ldots, t$,*[2]

$$\mathbb{E}_i\left[\exp\left(\theta\mathbf{E}_i\right)\right] \preceq \exp\left(g_t(\theta)\mathbf{U}_i\right) \quad \textit{almost surely for each } \theta \in \Theta_t. \tag{12}$$

Then, we have for any scalars $\eta \geq 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=0}^t \mathbf{E}_i\right\| \geq \eta \ \text{ and } \ \left\|\sum_{i=0}^t \mathbf{U}_i\right\| \leq \sigma^2\right) \leq 2d \cdot \inf_{\theta\in\Theta_t} \exp\left(-\theta\eta + g_t(\theta)\sigma^2\right).$$

The function g_t in [57, Theorem 2.3] is defined on the full positive set $(0, \infty)$, but the proof applies to any subset Θ_t. We use Lemma 1 to show the next result.

[2] The matrix exponential is defined by power series expansion: $\exp(\mathbf{A}) = \mathbf{I} + \sum_{i=1}^{\infty} \mathbf{A}^i/i!$.

Theorem 3 (Concentration of sub-exponential martingale difference) *Under Assumption 1, for any integer $t \geq 0$ and scalar $\eta \geq 0$, $\bar{\mathbf{E}}_t$ in (11) satisfies*

$$\mathbb{P}\left(\left\|\bar{\mathbf{E}}_t\right\| \geq \eta\right) \leq 2d \cdot \exp\left(-\frac{\eta^2/2}{\Upsilon_E^2 \sum_{i=0}^t z_{i,t}^2 + z_t^{(\max)} \Upsilon_E \eta}\right) \tag{13}$$

where $z_t^{(\max)} = \max_{i \in \{0,\ldots,t\}} z_{i,t}$.

Proof See Appendix C.1. □

The martingale concentration in Theorem 3 matches the matrix Bernstein results for independent noises $\{\mathbf{E}_i\}_{i=0}^t$ (cf. Theorems 6.1, 6.2 in [58]). For any $\delta \in (0, 1)$, if we let the right hand side of (13) be $\delta/(t+1)^2$, then we obtain that, with probability at least $1 - \delta/(t+1)^2$,

$$\left\|\bar{\mathbf{E}}_t\right\| \leq 8\Upsilon_E \sqrt{\log\left(\frac{d(t+1)}{\delta}\right)} \left(\sqrt{\sum_{i=0}^t z_{i,t}^2} \vee \sqrt{\log\left(\frac{d(t+1)}{\delta}\right)} \cdot z_t^{(\max)}\right). \tag{14}$$

We provide the following remark to discuss the fastest concentration rate.

Remark 2 To achieve the fastest concentration rate, we minimize the right hand side of (14) under the restriction $\sum_{i=0}^t z_{i,t} = 1$. Note that the minimum of both $\sum_{i=0}^t z_{i,t}^2$ and $z_t^{(\max)}$ is attained with equal weights, that is $z_{i,t} = 1/(t+1)$, for $i = 0, 1, \ldots, t$. Thus, the fastest concentration rate is attained with equal weights. Furthermore, a union bound over t leads to

$$\begin{aligned}
&\mathbb{P}\left(\forall t : \left\|\bar{\mathbf{E}}_t\right\| \leq 8\Upsilon_E \left(\sqrt{\frac{\log(d(t+1)/\delta)}{t+1}} \vee \frac{\log(d(t+1)/\delta)}{t+1}\right)\right) \\
&\geq 1 - \sum_{t=0}^{\infty} \frac{\delta}{(t+1)^2} = 1 - \frac{\pi^2 \delta}{6}.
\end{aligned} \tag{15}$$

We note that the square root term $\sqrt{\log(d(t+1)/\delta)/t+1}$ dominates the error bound for large t. Recalling from (11) that $z_{i,t} = (w_i - w_{i-1})/w_t$, we know $w_t = t+1$ for the equal weights. If fact, the concentration rate of $\|\bar{\mathbf{E}}_t\|$ relates to the superlinear convergence rate of $\mathbf{x}_t$ (see Theorem 1), because, as shown in the following sections, the convergence rate of $\mathbf{x}_t$ is proportional to $\|\bar{\mathbf{E}}_t\|$ when $\mathbf{x}_t$ is sufficiently close to $\mathbf{x}^\star$.

3 Convergence of uniform Hessian averaging

We now study the convergence of stochastic Newton with Hessian averaging. We consider the uniform averaging scheme, i.e., $w_t = t+1$, $\forall t \geq 0$. Our first result suggests that, with high probability, $\tilde{\mathbf{H}}_t \succ \mathbf{0}$ for all large t. This implies that the

Newton direction $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$ will be employed from some t onwards (cf. Line 6 of Algorithm 1). We recall that $\Upsilon = \Upsilon_E/\lambda_{\min}$, and e denotes the natural base.

Lemma 2 *Consider Algorithm 1 with $w_t = t+1$, $\forall t \geq 0$. Under Assumption 1, we let $\delta, \epsilon \in (0,1)$ with $d/\delta \geq e$. We also let*

$$\mathcal{T}_1 = 4\left(1 \vee (8\Upsilon/\epsilon)\right)^2 \log\left(d/\delta \cdot \{1 \vee (8\Upsilon/\epsilon)\}\right). \tag{16}$$

Then, with probability $1 - \delta\pi^2/6$, the event

$$\mathcal{E} = \bigcap_{t=\mathcal{T}_1}^{\infty} \left\{ \|\bar{\mathbf{E}}_t\| \leq 8\Upsilon_E \sqrt{\frac{\log(d(t+1)/\delta)}{t+1}} \right\} \tag{17}$$

occurs, which implies $(1-\epsilon)\lambda_{\min} \cdot \mathbf{I} \preceq \tilde{\mathbf{H}}_t \preceq (1+\epsilon)\lambda_{\max} \cdot \mathbf{I}$, $\forall t \geq \mathcal{T}_1$.

Proof See Appendix C.2. □

By Lemma 2, we initialize the convergence analysis from the iteration $t = \mathcal{T}_1$, and condition on the event $\mathcal{E}$. For $0 \leq t < \mathcal{T}_1$, the Newton system may or may not be solvable and the lower and upper bounds of $\tilde{\mathbf{H}}_t$ may or may not scale as $\lambda_{\min}$ and $\lambda_{\max}$ (cf. Line 6 of Algorithm 1). Thus, for the iterates $\mathbf{x}_{0:\mathcal{T}_1}$, we do not generally have guarantees on the convergence rate, but only know that the objective value is non-increasing, that is, $f(\mathbf{x}_0) \geq \cdots \geq f(\mathbf{x}_{\mathcal{T}_1})$.

We next provide a Q-linear convergence for the objective value $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ for $t \geq \mathcal{T}_1$.

Lemma 3 *Conditioning on the event (17), we let*

$$\phi = \frac{4\rho\beta(1-\beta)(1-\epsilon)}{\kappa^2(1+\epsilon)}$$

and have $f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \leq (1-\phi)(f(\mathbf{x}_t) - f(\mathbf{x}^\star))$, $\forall t \geq \mathcal{T}_1$, which implies R-linear convergence of the iterate error,

$$\|\mathbf{x}_t - \mathbf{x}^\star\| \leq \left\{ \frac{2}{\lambda_{\min}} (f(\mathbf{x}_0) - f(\mathbf{x}^\star))(1-\phi)^{t-\mathcal{T}_1} \right\}^{1/2}, \quad t \geq \mathcal{T}_1,$$
$$\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \left\{ 2\kappa (f(\mathbf{x}_0) - f(\mathbf{x}^\star))(1-\phi)^{t-\mathcal{T}_1} \right\}^{1/2}, \quad t \geq \mathcal{T}_1.$$

Proof See Appendix C.3. □

We next show that $\mathbf{x}_t$ stays in a neighborhood around $\mathbf{x}^\star$ for all large t. For this, we need a Lipschitz continuity condition.

Assumption 2 *(Lipschitz Hessian)* We assume $\mathbf{H}(\mathbf{x})$ is L-Lipschitz continuous. That is $\|\mathbf{H}(\mathbf{x}_1) - \mathbf{H}(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|$ for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$.

Combining Lemma 3 with Assumption 2 leads to the following corollary.

Corollary 1 *Consider Algorithm 1 with $w_t = t+1$, $\forall t \geq 0$. Under Assumptions 1 and 2, we let $\delta, \epsilon \in (0,1)$ with $d/\delta \geq e$, and define the neighborhood $\mathcal{N}_\nu$ as*

$$\mathcal{N}_\nu = \{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \nu \cdot \lambda_{\min}^{3/2}/L\} \quad \text{for } \nu \in (0,1]. \tag{18}$$

Then, with probability $1 - \delta\pi^2/6$, we have $\mathbf{x}_t \in \mathcal{N}_\nu$, for all $t \geq \mathcal{T}$ where $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2$ with $\mathcal{T}_1$ defined in (16) and

$$\mathcal{T}_2 = \frac{\kappa^2(1+\epsilon)}{4\rho\beta(1-\beta)(1-\epsilon)} \log\left(\frac{3L^2(f(\mathbf{x}_0) - f(\mathbf{x}^\star))}{\nu^2\lambda_{\min}^3}\right). \tag{19}$$

Proof See Appendix C.4. □

Combining (16) and (19), and using $O(\cdot)$ to neglect logarithmic factors and all constants except κ and Υ, we have $\mathcal{T} = O(\Upsilon^2 + \kappa^2)$ with high probability. Building on Corollary 1, we then show that the unit stepsize is accepted locally.

Lemma 4 *Under Assumption 2, suppose $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$. Then $\mu_t = 1$ if $\mathbf{x}_t \in \mathcal{N}_\nu$ and $(1-\psi)\mathbf{H}_t \preceq \tilde{\mathbf{H}}_t \preceq (1+\psi)\mathbf{H}_t$ with ν, ψ satisfying*

$$0 < \nu \leq \frac{2}{3}(1/2 - \beta), \qquad 0 < \psi \leq \frac{1/2-\beta}{3/2-\beta}. \tag{20}$$

Proof See Appendix C.5. □

The unit stepsize enables us to show a linear-quadratic error recursion.

Lemma 5 *Under Assumption 2 and suppose $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$, $\mathbf{x}_t \in \mathcal{N}_\nu$, and $(1-\psi)\mathbf{H}_t \preceq \tilde{\mathbf{H}}_t \preceq (1+\psi)\mathbf{H}_t$ with ν, ψ satisfying (20). Then, we have*

$$\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|_{\mathbf{H}^\star} \leq 3\left\{\frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}^2 + \|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\| \cdot \|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}\right\}.$$

Proof See Appendix C.6. □

Lemma 5 suggests that $\|\mathbf{x}_t - \mathbf{x}^\star\|$ exhibits local Q-linear convergence.

Corollary 2 *Under Assumption 2 and suppose $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$, $\mathbf{x}_t \in \mathcal{N}_\nu$, and $(1-\psi)\mathbf{H}_t \preceq \tilde{\mathbf{H}}_t \preceq (1+\psi)\mathbf{H}_t$ with ν, ψ satisfying (20). Then, we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 3(\nu+\psi)\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star},$$

which implies linear convergence provided $3(\nu+\psi) < 1$.

Given all the presented lemmas, we state the final convergence guarantee.

Theorem 4 *Consider Algorithm 1 with $w_t = t + 1$, $\forall t \geq 0$. Under Assumptions 1, 2, we let $\delta \in (0, 1)$ satisfy $d/\delta \geq e$, and let $\epsilon, \nu \in (0, 1)$ satisfy*

$$\epsilon \vee \nu \leq \frac{1}{3}\frac{0.5-\beta}{1.5-\beta} \wedge \frac{1}{48}. \tag{21}$$

Define the neighborhood $\mathcal{N}_\nu$ as in (18), and define $\mathcal{T} = \mathcal{T}_1 + \mathcal{T}_2$ with $\mathcal{T}_1$ given by (16) and $\mathcal{T}_2$ given by (19). We also let $\mathcal{J} = 4\mathcal{T}\kappa/\nu$. Then, with probability $1 - \delta\pi^2/6$, we have that $\mathbf{x}_{\mathcal{T}_1:\mathcal{T}+\mathcal{J}}$ converges R-linearly, $\mathbf{x}_{\mathcal{T}:\mathcal{T}+\mathcal{J}} \in \mathcal{N}_\nu$, and

$$\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 12\rho_t \|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t} - \mathbf{x}^\star\|_{\mathbf{H}^\star}, \quad \forall t \geq 0$$

with

$$\rho_t = \frac{4\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} + 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}}$$

satisfying $24\rho_t \leq 1$, $\forall t \geq 0$. □

Proof See Appendix C.7. □

We note from Theorem 4 that the condition on ϵ and ν does not depend on unknown quantities of objective function and noise level. The convergence rate ρ_t consists of two terms. The first term is due to the fact that $\mathcal{T} = O(\Upsilon^2 + \kappa^2)$ Hessians are accumulated in the global phase, and each of them contributes an error (in norm $\|\cdot\|_{\mathbf{H}^\star}$) as large as κ. Given these imprecise Hessians, the method cannot immediately converge superlinearly after $\mathcal{T}$ iterations. That is, $\rho_t \not\leq 1$ if $\mathcal{J} = t = 0$. We need $\mathcal{J} = O(\mathcal{T}\kappa)$ iterates to suppress the effect of these imprecise Hessians. The second term is due to the noise averaging, i.e., $\|\bar{\mathbf{E}}_t\|$, which decays slower than the first term. Thus, for sufficiently large t, the noise averaging will finally dominate the convergence rate.

We present the above observation in the following corollary. It suggests that the averaging scheme has three transition points; thus four convergence phases.

Corollary 3 *Under the setup of Theorem 4, Algorithm 1 has three transitions:*
(a): From $\mathbf{x}_0$ to $\mathbf{x}_\mathcal{T}$: the algorithm converges to a local neighborhood $\mathcal{N}_\nu$ from any initial point $\mathbf{x}_0$.
(b): From $\mathbf{x}_\mathcal{T}$ to $\mathbf{x}_{\mathcal{T}+\mathcal{J}}$: the sequence $\mathbf{x}_t$ stays in the neighborhood $\mathcal{N}_\nu$.
(Starting from $\mathbf{x}_{\mathcal{T}_1}$, the sequence $\mathbf{x}_t$ exhibits R-linear convergence)
(c): From $\mathbf{x}_{\mathcal{T}+\mathcal{J}}$ to $\mathbf{x}_{\mathcal{T}+\mathcal{J}+\mathcal{K}}$: the algorithm converges Q-superlinearly with

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 12\rho_t^{(1)}\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \quad \text{for} \quad \rho_t^{(1)} = \frac{8\mathcal{T}\kappa}{t+1},$$

where

$$0 \leq t - \mathcal{T} - \mathcal{J} \leq \mathcal{K}, \qquad \mathcal{K} := \frac{\mathcal{T}^2\kappa^2}{4\Upsilon^2\log(d\mathcal{T}/\delta)} - \mathcal{T} - \mathcal{J}.$$

(d): From $\mathbf{x}_{\mathcal{T}+\mathcal{J}+\mathcal{K}}$: the algorithm converges Q-superlinearly with

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 12\rho_t^{(2)}\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \quad \text{for} \quad \rho_t^{(2)} = 16\Upsilon\sqrt{\frac{\log(d(t+1)/\delta)}{t+1}},$$

where $t \geq \mathcal{T} + \mathcal{J} + \mathcal{K}$.

Proof See Appendix C.8. □

For an infinite iteration sequence $\{\mathbf{x}_t\}_{t=0}^{\infty}$ and with high probability, Corollary 3(a) suggests that the first transition is $\mathcal{T} = O(\kappa^2 + \Upsilon^2)$; Corollary 3(b) suggests that the second transition is $\mathcal{J} = O(\mathcal{T}\kappa)$; Corollary 3(c) suggests that the third transition is $\mathcal{K} = O(\mathcal{T}^2\kappa^2/\Upsilon^2)$; and Corollary 3(d) suggests that the final rate is $\rho_t^{(2)} = O(\Upsilon\sqrt{\log t/t})$. This recovers Theorem 1. Recalling that Υ typically does not exceed κ, in this case, we have $\mathcal{T} = O(\kappa^2)$, $\mathcal{J} = O(\kappa^3)$, and $\mathcal{K} = O(\kappa^6/\Upsilon^2)$. This suggests a trade-off between the final superlinear rate and the final transition. When the oracle noise level Υ is small, a faster superlinear rate is finally attained, but $\mathcal{K}$ also increases, meaning that the time to attain the final rate is further delayed. By Examples 1 and 2, Υ decays as sample/sketch size increases. Thus, the final superlinear rate is improved by increasing the sample/sketch size s, however the effect of this change may be delayed due to the rate/transition trade-off. Fortunately, as we will see in the following section, this trade-off can be optimized via *weighted* Hessian averaging.

4 Convergence of averaging with general weights

Although the superlinear convergence of stochastic Newton with uniform Hessian averaging (Corollary 3) is promising, since the scheme eventually attains the optimal superlinear rate implied by the central limit theorem, a clear drawback is the delayed transition—the scheme spends quite a long time before attaining the final rate. In this section, we study the relationship between transitions and general weight sequences. We consider performing Algorithm 1 with a weight sequence w_t that satisfies the following general condition.

Assumption 3 We assume $w_t = w(t)$ for all integer $t \geq 0$, where $w(\cdot) : \mathbb{R} \to \mathbb{R}$ is a real function satisfying (i) $w(\cdot)$ is twice differentiable; (ii) $w(-1) = 0, w(t) > 0, \forall t \geq 0$; (iii) $w'(-1) \geq 0$; (iv) $w''(t) \geq 0, \forall t \geq -1$; (v) $w(t+1)/w(t) \vee w'(t+1)/w'(t) \leq \Psi$, $\forall t \geq 0$ for some $\Psi \geq 1$.

By the above assumption, we specialize the result of noise averaging in (14) as follows.

Lemma 6 *Under Assumptions 1 and 3, for any* $t \geq 0$, *with probability* $1 - \delta/(t+1)^2$,

$$\|\bar{\mathbf{E}}_t\| \leq 8\Psi\Upsilon_E \left(\sqrt{\log\left(\frac{d(t+1)}{\delta}\right)\frac{w'(t)}{w(t)}} \vee \log\left(\frac{d(t+1)}{\delta}\right)\frac{w'(t)}{w(t)} \right). \tag{22}$$

Proof See Appendix C.9. □

Naturally, to have $\|\bar{\mathbf{E}}_t\|$ concentrate, we require $\lim_{t\to\infty} \log(\frac{d(t+1)}{\delta})\frac{w'(t)}{w(t)} = 0$. It is easy to see that for some weight sequences, such as $w(t) = \exp(t) - \exp(-1)$, such a requirement cannot be satisfied, which makes convergence fail. On the other

hand, this is reasonable since $z_{i,t} \propto z_i = w_i - w_{i-1} = \exp(i) - \exp(i-1) = (1-1/e)\exp(i)$, which means that we assign an exponentially large weight to the current Hessian estimate. Such an assignment preserves recent Hessian information better, but it diminishes the previous estimates too quickly to let the noise averaging concentrate.

Given the above concentration results, we have a similar result to Lemma 2.

Lemma 7 *Consider Algorithm 1 with w_t satisfying Assumption 3. Under Assumption 1, for any $\delta, \epsilon \in (0,1)$, we let*

$$\mathcal{I}_1 := \mathcal{I}_1(\epsilon, \delta) = \sup_t \left\{ t : \log\left(\frac{d(t+1)}{\delta}\right) \frac{w'(t)}{w(t)} \geq \left(\frac{\epsilon}{8\Psi\Upsilon} \wedge 1\right)^2 \right\} + 1. \tag{23}$$

Then, with probability $1 - \delta\pi^2/6$, the event

$$\mathcal{E} = \bigcap_{t=\mathcal{I}_1}^{\infty} \left\{ \|\bar{\mathbf{E}}_t\| \leq 8\Psi\Upsilon_E \sqrt{\log\left(\frac{d(t+1)}{\delta}\right) \frac{w'(t)}{w(t)}} \right\} \tag{24}$$

occurs, which implies $(1-\epsilon)\lambda_{\min} \cdot \mathbf{I} \preceq \tilde{\mathbf{H}}_t \preceq (1+\epsilon)\lambda_{\max} \cdot \mathbf{I}$, $\forall t \geq \mathcal{I}_1$.

Proof See Appendix C.10. □

We note that Lemma 3 and Corollary 1 still hold for general weight sequences. Thus, we let

$$\begin{aligned} \mathcal{I} := \mathcal{I}(\epsilon, \delta, \nu) &= \mathcal{I}_1(\epsilon, \delta) + \mathcal{T}_2 \\ &\overset{(19)}{=} \mathcal{I}_1(\epsilon, \delta) + \frac{\kappa^2(1+\epsilon)}{4\rho\beta(1-\beta)(1-\epsilon)} \log\left(\frac{3L^2(f(\mathbf{x}_0) - f(\mathbf{x}^\star))}{\nu^2\lambda_{\min}^3}\right), \end{aligned} \tag{25}$$

and know that $\mathbf{x}_t \in \mathcal{N}_\nu$ for all $t \geq \mathcal{I}$. Lemmas 4, 5 and Corollary 2 also carry over to the setting of general weight sequences. Building on these results, we state the final convergence guarantee.

Theorem 5 *Consider Algorithm 1 with w_t satisfying Assumption 3. Under Assumptions 1, 2, we let $\delta, \epsilon, \nu \in (0,1)$ and ϵ, ν satisfy*

$$\epsilon \vee \nu \leq \frac{1}{5}\frac{0.5-\beta}{1.5-\beta} \wedge \frac{1}{48\Psi}. \tag{26}$$

Define the neighborhood $\mathcal{N}_\nu$ as in (18), and define $\mathcal{I}$ as in (25). We also let $\mathcal{U}$ be $w(\mathcal{I}+\mathcal{U}) = 2w(\mathcal{I}-1)\kappa/\nu$. Then, with probability $1-\delta\pi^2/6$, we have that $\mathbf{x}_{\mathcal{I}_1:\mathcal{I}+\mathcal{U}}$ converges R-linearly, $\mathbf{x}_{\mathcal{I}:\mathcal{I}+\mathcal{U}} \in \mathcal{N}_\nu$, and

$$\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 6\theta_t \|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t} - \mathbf{x}^\star\|_{\mathbf{H}^\star}, \quad \forall t \geq 0,$$

with

$$\theta_t = \frac{6w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U}+t)} + 8\Psi\Upsilon\sqrt{\log\left(\frac{d(\mathcal{I}+\mathcal{U}+t+1)}{\delta}\right)\frac{w'(\mathcal{I}+\mathcal{U}+t)}{w(\mathcal{I}+\mathcal{U}+t)}}$$

satisfying $12\Psi\theta_t \leq 1$, $\forall t \geq 0$.

Proof See Appendix C.11. □

The proof of Theorem 5 is more involved than the one of Theorem 4. In particular, when dealing with a critical term $\sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}+t}(w(j)-w(j-1))\|\mathbf{x}_j-\mathbf{x}^\star\|_{\mathbf{H}^\star}$, the analysis of Theorem 4 uses the facts that $w(j)-w(j-1)=1$ and $\mathbf{x}_j$ converges linearly (cf. Corollary 2). However, that analysis does not apply for general weights, since plugging the setup $w(j)=j+1$ masks some critical properties of $w(\cdot)$, such as $w'(j)\geq 0$ and $w(j+1)/w(j)\leq\Psi$, $\forall j\geq 0$ ($\Psi=2$ for $w(j)=j+1$). In contrast, for Theorem 5, we separate the above term into two terms, $\sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}}(w(j)-w(j-1))\|\mathbf{x}_j-\mathbf{x}^\star\|_{\mathbf{H}^\star}$ and $\sum_{j=\mathcal{I}+\mathcal{U}+1}^{\mathcal{I}+\mathcal{U}+t}(w(j)-w(j-1))\|\mathbf{x}_j-\mathbf{x}^\star\|_{\mathbf{H}^\star}$. The first term is bounded by $w(\mathcal{I}+\mathcal{U})\nu$ since $\mathbf{x}_j\in\mathcal{N}_\nu$ for $\mathcal{I}\leq j\leq\mathcal{I}+\mathcal{U}$. The second term, which is proven to have the same bound as the first term, not only utilizes the linear convergence of $\mathbf{x}_j$ *with a rate depending on* Ψ (proved by induction), but also utilizes general properties of $w(\cdot)$ in Assumption 3.

We arrive at the following corollary for the iteration transitions. The proof is the same as for Corollary 3, and is omitted.

Corollary 4 *Under the setup of Theorem 5, Algorithm 1 has three transitions:*
(a): From $\mathbf{x}_0$ *to* $\mathbf{x}_\mathcal{I}$*: the algorithm converges to a local neighborhood* $\mathcal{N}_\nu$ *from any initial point* $\mathbf{x}_0$.
(b): From $\mathbf{x}_\mathcal{I}$ *to* $\mathbf{x}_{\mathcal{I}+\mathcal{U}}$*: the sequence* $\mathbf{x}_t$ *stays in the neighborhood* $\mathcal{N}_\nu$.
(Starting from $\mathbf{x}_{\mathcal{I}_1}$*, the sequence* $\mathbf{x}_t$ *exhibits R-linear convergence)*
(c): From $\mathbf{x}_{\mathcal{I}+\mathcal{U}}$ *to* $\mathbf{x}_{\mathcal{I}+\mathcal{U}+\mathcal{V}}$*: the algorithm converges Q-superlinearly with*

$$\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\leq 6\theta_t^{(1)}\|\mathbf{x}_t-\mathbf{x}^\star\|_{\mathbf{H}^\star}\quad \textit{for}\quad \theta_t^{(1)}:=\frac{14w(\mathcal{I}-1)\kappa}{w(t)},$$

where

$$0\leq t-\mathcal{I}-\mathcal{U}\leq\mathcal{V},\qquad \mathcal{V}:=\operatorname*{argmin}_{t\geq\mathcal{I}+\mathcal{U}}\left\{w(t)w'(t)\log\left(\frac{d(t+1)}{\delta}\right)\geq\frac{w(\mathcal{I}-1)^2\kappa^2}{\Psi^2\Upsilon^2}\right\}.$$

(d): From $\mathbf{x}_{\mathcal{I}+\mathcal{U}+\mathcal{V}}$*: the algorithm converges Q-superlinearly with*

$$\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\leq 6\theta_t^{(2)}\|\mathbf{x}_t-\mathbf{x}^\star\|_{\mathbf{H}^\star}\ \ \textit{for}\ \ \theta_t^{(2)}=14\Psi\Upsilon\sqrt{\frac{w'(t)}{w(t)}\log\left(\frac{d(t+1)}{\delta}\right)},$$

where $t\geq\mathcal{I}+\mathcal{U}+\mathcal{V}$.

Table 2 Comparison of different averaging schemes, in terms of how many iterations it takes to transition to each superlinear phase, and the convergence rates achieved

Weights w_t	Initial superlinear phase		Final superlinear phase	
	$\mathcal{U}$	rate $\theta_t^{(1)}$	$\mathcal{V}$	rate $\theta_t^{(2)}$
$t+1$ (Ex. 3)	κ^3	κ^3/t	κ^6/Υ^2	$\Upsilon\sqrt{\log(t)/t}$
$(t+1)^p$ (Ex. 4)	$\kappa^{2+1/p}$	κ^{2p+1}/t^p	$\kappa^{\frac{4p+2}{2p-1}}/(p\Upsilon^2)^{\frac{1}{2p-1}}$	$\Upsilon\sqrt{p\log(t)/t}$
$(t+1)^{\log(t+1)}$ (Ex. 5)	κ^2	$\kappa^{4\log(\kappa)+1}/t^{\log(t)}$	κ^2	$\Upsilon\log(t)/\sqrt{t}$

We drop constant factors as well as logarithmic dependence on d and $1/\delta$, and assume that $1/\text{poly}(\kappa) \leq \Upsilon \leq O(\kappa)$

Given Corollary 4, we provide the following examples. Again, we use $O(\cdot)$ to neglect the logarithmic factors and the dependence on all constants except κ and Υ. We emphasize that for an iteration sequence, the three transition points occur only with high probability. For sake of presentation, we will not repeat "with high probability" for each $O(\cdot)$. As mentioned, typically, $\Upsilon = O(\kappa)$ in practice. We note that for all considered w_t sequences, Assumption 3 is satisfied and $\mathcal{I}_1$ has the same order as $\mathcal{T}_1$ (cf. (16)), so that $\mathcal{I} = O(\mathcal{T}) = O(\Upsilon^2 + \kappa^2)$. In other words, the weighted averaging and uniform averaging take the same order of iterations to get into the same local neighborhood. In the following examples, we show how the second and third transition points $\mathcal{U}$ and $\mathcal{V}$, as well as the superlinear rates $\theta_t^{(1)}$ and $\theta_t^{(2)}$, change for different weight sequences (see Table 2 for a summary of the case $\Upsilon \leq O(\kappa)$, i.e. $\mathcal{I} = O(\kappa^2)$).

Example 3 (*Uniform averaging*) Let $w_t = t+1$. Then, the superlinear rates are:

$$\theta_t^{(1)} = O\left(\frac{\mathcal{I}\kappa}{t}\right), \qquad \theta_t^{(2)} = O\left(\Upsilon\sqrt{\frac{\log(dt/\delta)}{t}}\right).$$

Moreover, the second and third transition points are given by $\mathcal{U} = O(\mathcal{I}\cdot\kappa)$ and $\mathcal{V} = O(\mathcal{I}^2\kappa^2/\Upsilon^2)$. The above example recovers Corollary 3 (up to constant scaling factors). It achieves the best final rate $\theta_t^{(2)}$ ensured by the central limit theorem, but the initial rate $\theta_t^{(1)}$ is sub-optimal. Thus, this scheme takes a long time for the iterates to reach the final rate.

Example 4 (*Accelerated initial rate*) Let $w_t = (t+1)^p$ for $p \geq 1$. The superlinear rates are:

$$\theta_t^{(1)} = O\left(\frac{\mathcal{I}^p\kappa}{t^p}\right), \qquad \theta_t^{(2)} = O\left(\Upsilon\sqrt{p\cdot\frac{\log(dt/\delta)}{t}}\right).$$

Moreover, the second and third transition points are given by

$$(\mathcal{I}+\mathcal{U})^p = O(\mathcal{I}^p\kappa) \Longleftarrow \mathcal{U} = O(\mathcal{I}\cdot\kappa^{1/p}),$$

$$p\mathcal{V}^{2p-1}\log(d\mathcal{V}/\delta) \geq \mathcal{I}^{2p}\kappa^2/\Upsilon^2 \Longleftarrow \mathcal{V} = O\left(\mathcal{I}^{\frac{2p}{2p-1}} \cdot \left\{\kappa^2/(p\Upsilon^2)\right\}^{\frac{1}{2p-1}}\right).$$

In the above example, we substantially accelerate the initial superlinear rate $\theta_t^{(1)}$, while preserving the final rate $\theta_t^{(2)}$ up to a constant factor p. We observe that the transitions $\mathcal{U}$ and $\mathcal{V}$ are both improved upon Example 3. In particular, the dependence of κ in $\mathcal{U}$ is improved from κ to $\kappa^{1/p}$, and the dependence of $\mathcal{I}$ in $\mathcal{V}$ is improved from $\mathcal{I}^2$ to $\mathcal{I}^{2p/(2p-1)}$. An important observation is that both $\mathcal{U}$ and $\mathcal{V}$ contract to $\mathcal{I}$ as p goes to ∞, which inspires us to consider the following sequence.

Example 5 (*Optimal transition points*) Let $w_t = (t+1)^{\log(t+1)}$. Then, the superlinear rates are:

$$\theta_t^{(1)} = O\left(\frac{\mathcal{I}^{\log(\mathcal{I})}\kappa}{t^{\log(t)}}\right), \qquad \theta_t^{(2)} = O\left(\Upsilon\sqrt{\log(t)\frac{\log(dt/\delta)}{t}}\right),$$

We now derive the second and third transition points. In particular, $\mathcal{U}$ can be obtained from:

$$\begin{aligned}&(\mathcal{I}+\mathcal{U})^{\log(\mathcal{I}+\mathcal{U})} = O(\mathcal{I}^{\log\mathcal{I}}\kappa)\\ &\Longleftarrow \mathcal{I}+\mathcal{U} = O\left(\exp\left(\sqrt{\log\left(\mathcal{I}^{\log\mathcal{I}}\kappa\right)}\right)\right) = O(\mathcal{I}) \Longleftarrow \mathcal{U} = O(\mathcal{I}),\end{aligned}$$

where the first implication uses the fact $y = x^{\log x} \Longleftrightarrow x = \exp(\sqrt{\log y})$, as well as the fact $\log\kappa \leq (\log\mathcal{I})^2$. The third transition $\mathcal{V}$ can be obtained from:

$$\begin{aligned}&\mathcal{V}^{2\log\mathcal{V}-1}\log\mathcal{V}\log d\mathcal{V}/\delta \geq \mathcal{I}^{2\log\mathcal{I}}\kappa^2/\Upsilon^2\\ &\Longleftarrow (2\log\mathcal{V}-1)\log\mathcal{V} \geq (2\log\mathcal{I}+1)\log\mathcal{I} + 2\log\left(\frac{1}{\Upsilon}\right),\end{aligned}$$

where the implication uses $\kappa^2 \leq \mathcal{I}$. To let the right hand side hold, we let $\mathcal{V} = \xi\cdot\mathcal{I}$ and require

$$\begin{aligned}&((2\log\xi - 2) + 2\log\mathcal{I} + 1)(\log\mathcal{I} + \log\xi) \geq (2\log\mathcal{I}+1)\log\mathcal{I} + 2\log\left(\frac{1}{\Upsilon}\right)\\ &\Longleftarrow \begin{cases}2\log\xi - 2 \geq 1\\ \log\xi\log\kappa \geq \log(1/\Upsilon)\end{cases} \Longleftarrow \xi \geq \exp\left(1.5 \vee \log(1/\Upsilon)/\log\kappa\right).\end{aligned}$$

Thus, the final transition point $\mathcal{V}$ can be chosen as

$$\mathcal{V} = O\left(\mathcal{I}\cdot\exp\left(1 + \frac{\log 1/\Upsilon}{\log\kappa}\right)\right).$$

Note that, as long as the oracle noise Υ is bounded below as $\Upsilon \geq 1/\mathrm{poly}(\kappa)$, we have $\mathcal{V} = O(\mathcal{I})$. Since, as we mentioned, in practice the oracle noise actually grows

proportionally with κ, this is an extremely mild condition. However, it is technically necessary, since, if the Hessian oracle always returned the true Hessian, i.e., $\Upsilon = 0$, then $\theta_t^{(2)} = 0$, so we would necessarily have $\mathcal{V} = \infty$ (i.e., Hessian averaging is not helpful when we have the exact Hessian). In the above example, all of the transition points $\mathcal{I}, \mathcal{U}, \mathcal{V}$ are within constant factors of one another (not even logarithmic factors), while we sacrifice only $\sqrt{\log(t)}$ in the final superlinear rate. The above results recover Theorem 2. This suggests that, using the weight sequence $w_t = (t+1)^{\log(t+1)}$, the averaging scheme transitions from the global phase to the local superlinearly convergent phase smoothly, and the local neighborhood that separates the two phases is consistent with the neighborhood of classical Newton methods that separate the global phase (i.e., the damped Newton phase) and the local quadratically convergent phase (cf. Lemma 5 with $\tilde{\mathbf{H}}_t = \mathbf{H}_t$).

5 Numerical experiments

We implement the proposed Hessian averaging schemes and compare them with popular baselines including BFGS, subsampled Newton methods and sketched Newton methods[3]. We focus on the regularized logistic regression problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-b_i \mathbf{a}_i^\top \mathbf{x}\right)\right) + \frac{\nu}{2}\|\mathbf{x}\|^2,$$

where $\{(b_i, \mathbf{a}_i)\}_{i=1}^n$ with $b_i \in \{-1, 1\}$ are input-label pairs. We let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)^\top \in \mathbb{R}^{n\times d}$ be the data matrix.

Data generating process: We generate several data matrices with $(n, d, \nu) = (1000, 100, 10^{-3})$, varying their properties. First, we vary the coherence of $\mathbf{A}$, since higher coherence leads to higher variance for subsampled Hessian estimates, compared to sketched Hessian estimates. We use $\tau_{\mathbf{A}}$ to denote the coherence of $\mathbf{A}$, which is defined as

$$\tau_{\mathbf{A}} = \frac{n}{d} \cdot \max_{i=1,\ldots,n} \|\mathbf{U}_{i,\cdot}\|^2$$

where $\mathbf{A} = \mathbf{U}\acute{\mathbf{V}}^\top$ is the reduced singular value decomposition of $\mathbf{A}$ with $\mathbf{U} \in \mathbb{R}^{n\times d}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$. Second, we vary the condition number $\kappa_{\mathbf{A}}$ of $\mathbf{A}$, since (as suggested by our theory), higher condition number leads to slower convergence and delayed transitions to superlinear rate. The detailed generalization of $\mathbf{A}$ is as follows. We fix $\mathbf{V} = \mathbf{I} \in \mathbb{R}^{d\times d}$. For low coherence, we generate a $n \times d$ random matrix with each entry being independently generated from $\mathcal{N}(0, 1)$, and let $\mathbf{U}$ be the left singular vectors of that matrix. For high coherence, we divide each row of $\mathbf{U}$ by $\sqrt{\mathbf{z}_i}$, where $\mathbf{z}_i$ is independently sampled from Gamma distribution with shape 0.5 and scale 2. We observe that the coherence of the low coherence matrix is ≈ 1 (minimal), while the coherence of the high coherence matrix is $\approx 10 = n/d$ (maximal). For either low or high coherence, we vary $\kappa_{\mathbf{A}} = \{d^{0.5}, d, d^{1.5}\}$. For each $\kappa_{\mathbf{A}}$, we let singular values

[3] All results can be reproduced via https://github.com/senna1128/Hessian-Averaging.

in $''$ vary from 1 to $\kappa_{\mathbf{A}}$ with equal spacing, and finally form the matrix $\mathbf{A} = \mathbf{U}''$. We also generate a random vector $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, 1/d \cdot \mathbf{I})$, and the response $b_i \in \{-1, 1\}$ with $P(b_i = 1) = 1/(1 + \exp(-\mathbf{a}_i^\top \mathbf{x}))$, $\forall i = 1, \ldots, n$.

Methods: We implement the deterministic BFGS method, and stochastic subsampled and sketched Newton methods. Given a batch size s, the subsampled Newton method computes the Hessian as

$$\hat{\mathbf{H}}(\mathbf{x}) = \frac{1}{s} \sum_{j \in \xi} l_j \cdot \mathbf{a}_j \mathbf{a}_j^\top + \nu \cdot \mathbf{I} \quad \text{with} \quad l_j = \frac{\exp(-b_j \mathbf{a}_j^\top \mathbf{x})}{\left(1 + \exp(-b_j \mathbf{a}_j^\top \mathbf{x})\right)^2},$$

where the index set ξ has size s and is generated uniformly from $[1, n]$ without replacement. The sketched Newton method instead computes the Hessian as

$$\hat{\mathbf{H}}(\mathbf{x}) = \mathbf{A}^\top \mathbf{D}^{1/2} \mathbf{S}^\top \mathbf{S} \mathbf{D}^{1/2} \mathbf{A} + \nu \cdot \mathbf{I} \quad \text{with} \quad \mathbf{D} = \operatorname{diag}(l_1, \ldots, l_n),$$

where $\mathbf{S} \in \mathbb{R}^{s \times n}$ is the sketch matrix. We consider three sketching methods, one dense (slow and accurate) and two sparse variants (fast but less accurate).

(i) Gaussian sketch: we let $\mathbf{S}_{i,j} \overset{iid}{\sim} \mathcal{N}(0, 1)$.
(ii) CountSketch [13]: for each column of $\mathbf{S}$, we randomly pick a row and realize a Rademacher variable.
(iii) LESS-Uniform [19]: for each row of $\mathbf{S}$, we randomly pick a fixed number of nonzero entries and realize a Rademacher variable in each nonzero entry. We let the number of nonzero entries in each row to be $0.1d$.
For (i)–(iii), the sketches are scaled appropriately so that $\mathbb{E}[\hat{\mathbf{H}}(\mathbf{x})] = \mathbf{H}(\mathbf{x})$.

For each of the above Hessian oracles (both sketched and subsampled), we consider three variants of the stochastic Newton methods, depending on how the final Hessian estimate is constructed.
(1) NoAvg: the standard method that uses the oracle estimate directly;
(2) UnifAvg: the uniform Hessian averaging (i.e., $w_t = t + 1$ from Theorem 1);
(3) WeightAvg: the universal weighted Hessian averaging (i.e., $w_t = (t+1)^{\log(t+1)}$ from Theorem 2).

The overall results are summarized in Table 3, where we show the number of iterations until convergence for each experiment (we use $\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 10^{-6}$ as the criterion). We display the median over 50 independent runs. The first general takeaway is that our proposed WeightAvg performs best overall, and it is the most robust to different problem settings, when varying the coherence $\tau_{\mathbf{A}}$, condition number $\kappa_{\mathbf{A}}$, sample size s, and sketch/sample type. In particular, among the three variants of NoAvg/UnifAvg/WeightAvg, the latter is the only one to beat BFGS in all settings with sample size as small as $s = 0.5d$, as well as the only one to successfully converge within the iteration budget (999 iterations) for all Hessian oracles. Nevertheless, there are a few problem instances where either NoAvg or UnifAvg perform somewhat better than WeightAvg, which can also be justified by our theory. In the cases where NoAvg performs best, the oracle noise is particularly small (due to the use of a dense Gaussian sketch and/or a large sketch size), which means that averaging out the noise is not helpful until long after the method has converged very close to the optimum. On the other

Table 3 We show the median over 50 runs of the number of iterations until convergence, i.e., until $\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 10^{-6}$, where "—" indicates exceeding 999 iterations

Setup			Newton Sketch			Subsampled	BFGS
$\tau_{\mathbf{A}}$	$\kappa_{\mathbf{A}}$	s	Gaussian	CountSketch	LESS-uniform	Newton	
1	$d^{0.5}$	$0.25d$	67/**29**/35	67/**29**/35	66/**29**/35	67/**29**/36	177
		$0.5d$	53/**22**/25	53/**22**/25	53/**21**/25	53/**22**/25	
		d	38/**17**/19	38/**17**/18	37/**17**/19	38/**16**/18	
		$5d$	16/**11**/**11**	16/**11**/**11**	16/**11**/12	12/**9**/**9**	
	d	$0.25d$	—/**41**/52	—/**41**/52	—/**41**/53	—/**45**/60	219
		$0.5d$	—/**31**/34	—/**31**/35	—/**31**/35	—/**34**/39	
		d	244/**24**/**24**	246/**24**/**24**	243/**24**/**24**	315/**26**/**26**	
		$5d$	20/16/**14**	20/17/**14**	20/16/**14**	18/15/**13**	
	$d^{1.5}$	$0.25d$	—/—/**80**	—/—/**81**	—/—/**98**	—/—/**371**	295
		$0.5d$	—/—/**67**	—/—/**66**	—/—/**72**	—/—/**217**	
		d	270/—/**59**	—/—/**59**	—/—/**59**	—/—/**122**	
		$5d$	**27**/—/54	97/—/**54**	**38**/—/54	—/—/**54**	
10	$d^{0.5}$	$0.25d$	60/**29**/34	60/**29**/33	60/**33**/37	57/60/**51**	178
		$0.5d$	48/**22**/24	47/**22**/24	46/**25**/26	45/45/**41**	
		d	38/**17**/18	35/**17**/18	34/**20**/**20**	34/36/**33**	
		$5d$	15/**11**/**11**	16/**11**/**11**	16/**12**/**12**	26/17/**15**	
	d	$0.25d$	—/91/**52**	—/90/**51**	—/117/**63**	—/242/**147**	252
		$0.5d$	—/74/**35**	—/74/**36**	—/93/**44**	—/164/**106**	
		d	101/65/**27**	121/63/**27**	151/74/**31**	308/114/**69**	
		$5d$	21/51/**19**	27/52/**18**	25/55/**19**	77/74/**25**	
	$d^{1.5}$	$0.25d$	—/538/**80**	—/549/**78**	—/754/**110**	—/—/**294**	273
		$0.5d$	—/518/**62**	—/499/**63**	—/575/**77**	—/946/**187**	
		d	235/475/**51**	429/489/**53**	729/510/**59**	—/659/**121**	
		$5d$	**35**/431/43	53/422/**43**	**42**/438/43	321/462/**51**	

In the case of stochastic methods, we provide three numbers for (the best one in bold): NoAvg, UnifAvg, and WeightAvg, which correspond to the standard method (without Hessian averaging), the method with uniform averaging ($w_t = t + 1$, as in Theorem 1), and the method with weighted averaging ($w_t = (t + 1)^{\log(t+1)}$, as in Theorem 2), respectively. For the setup, we use $\tau_{\mathbf{A}}$ to denote the coherence number of $\mathbf{A}$; $\kappa_{\mathbf{A}}$ to denote the condition number of $\mathbf{A}$; and s to denote the sample/sketch size for the stochastic methods

hand, the cases where UnifAvg performs best are characterized by a well-conditioned objective function, where the superlinear phase of the optimization is reached almost instantly, so the slightly weaker superlinear rate of WeightAvg compared to UnifAvg manifests itself before reaching convergence (i.e., the additional factor of $\sqrt{\log(t)}$ in Theorem 2).

To investigate the performance of Hessian averaging more closely, we present selected convergence plots in Fig. 1. The figure shows decay of the error in the log-scale, so that a linear slope indicates linear convergence whereas a concave slope implies superlinear rate. Here, we used subsampling as the Hessian oracle, varying

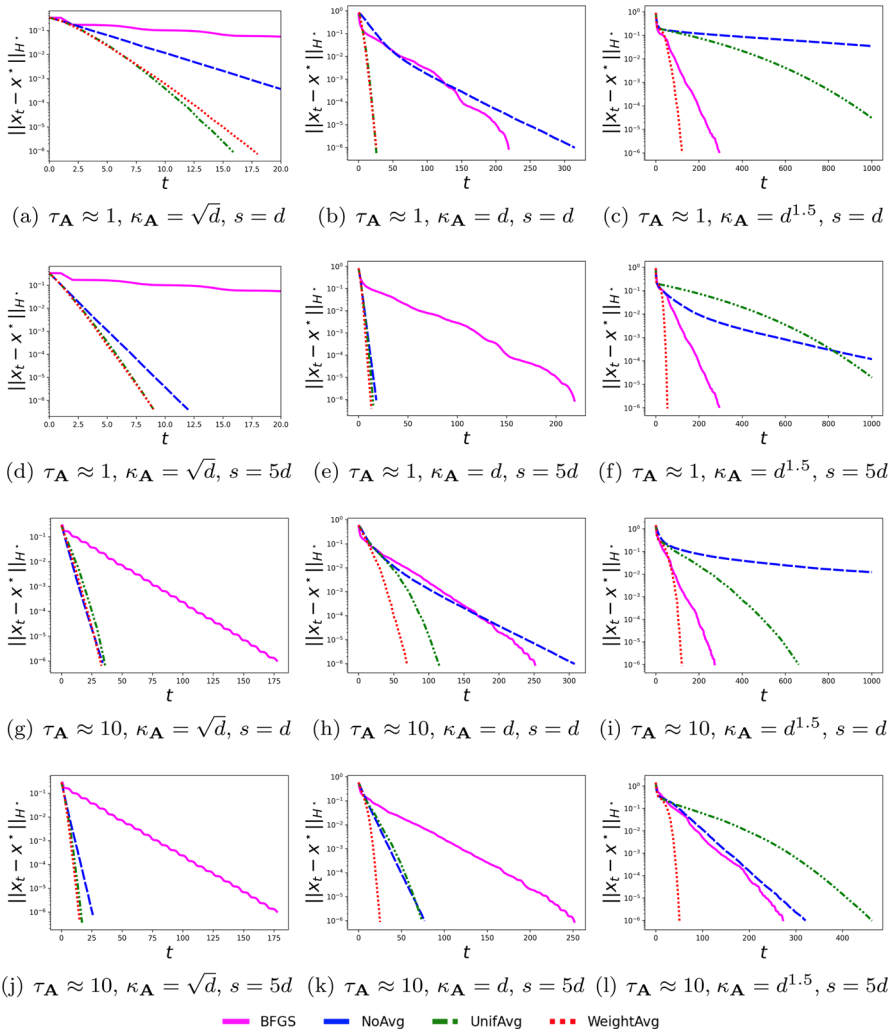

Fig. 1 Convergence plots for Subsampled Newton with several Hessian averaging schemes, compared to the standard method without averaging (NoAvg) and to BFGS. For all of the plots, we have $(n, d, \nu) = (1000, 100, 10^{-3})$. We vary the data coherence $\tau_{\mathbf{A}}$, condition number $\kappa_{\mathbf{A}}$ and the subsample size s

the coherence $\tau_{\mathbf{A}}$, the condition number $\kappa_{\mathbf{A}}$, and sample size s, and compared the Hessian averaging schemes UnifAvg and WeightAvg against the baselines of standard Subsampled Newton (i.e., NoAvg) and BFGS. We make the following observations:

(a) Subsampled Newton with Hessian averaging (UnifAvg and WeightAvg) exhibits a clear superlinear rate, observable in how its error plot curves away from the linear convergence of NoAvg. We note that BFGS also exhibits a superlinear rate, but only much later in the convergence process.
(b) The gain from Hessian averaging (relative to NoAvg) is more significant both for highly ill-conditioned problems (large condition number κ) and for noisy Hessian

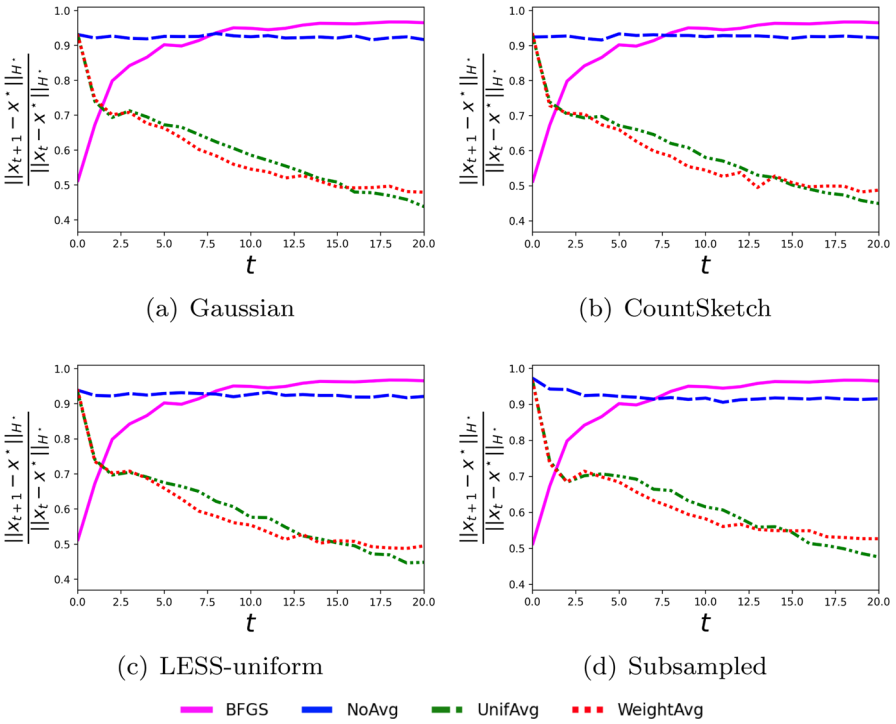

Fig. 2 Convergence rates for different Hessian oracles and averaging schemes, with $(\tau_{\mathbf{A}}, \kappa_{\mathbf{A}}, s) = (1, d, d)$. We truncate iterations at 20 to highlight the superlinear rate of UnifAvg and WeightAvg

oracles (small sample size s). For example, in the setting of $(\kappa, s) = (d^{1.5}, d)$ for both low and high coherence, standard Subsampled Newton (NoAvg) converges orders of magnitude slower than BFGS, and yet after introducing weighted averaging (WeightAvg), it beats BFGS by a factor of at least 2.5.

(c) For small condition number, the two Hessian averaging schemes (UnifAvg and WeightAvg) perform similarly, although in the setting of $(\kappa, s) = (\sqrt{d}, d)$, the superlinear rate of UnifAvg is slightly better than that of WeightAvg (cf. Fig. 1a), which aligns with a slightly better rate in Theorem 1 compared to Theorem 2.

(d) For highly ill-conditioned problems, WeightAvg converges much faster than UnifAvg. This is because, as suggested by Theorem 1, it takes much longer for UnifAvg to transition to its fast rate. In fact, in the settings of $(\tau_{\mathbf{A}}, \kappa_{\mathbf{A}}, s) = (1, d^{1.5}, 5d)$ and $(\tau_{\mathbf{A}}, \kappa_{\mathbf{A}}, s) = (10, d, 5d)$, UnifAvg initially trails behind NoAvg, which is a consequence of the distortion of the averaged Hessian by the noisy estimates from the early global convergence.

We next investigate the convergence rate of different methods, and aim to show that Hessian averaging leads to Q-superlinear convergence, while subsampled/sketched Newton (with fixed sample/sketch size) only exhibit Q-linear convergence. Figure 2 plots $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} / \|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}$ versus t. From the figure, we indeed observe that our method with different weight sequences (UnifAvg and WeightAvg) always

exhibits a Q-superlinear convergence, and the superlinear rate exhibits the trend of $(1/t)^{t/2}$ matching the theory up to logarithmic factors. We also observe that subsampled/sketched Newton methods with different sketch matrices (NoAvg) always exhibit a Q-linear convergence. These observations are all consistent with our theory.

6 Conclusions

This paper investigated the stochastic Newton method with Hessian averaging. In each iteration, we obtain a stochastic Hessian estimate and then average it with all the past Hessian estimates. We proved that the proposed method exhibits a local Q-superlinear convergence rate, although the non-asymptotic rate and the transition points are different for different weight sequences. In particular, we proved that uniform Hessian averaging finally achieves $(\Upsilon\sqrt{\log t/t})^t$ superlinear rate, which is faster than the other weight sequences, but it may take as many as $O(\Upsilon^2+\kappa^6/\Upsilon^2)$ iterations with high probability to get to this rate. We also observe that using weighted averaging $w_t = (t+1)^{\log(t+1)}$, the averaging scheme transitions from the global phase to the superlinear phase smoothly after $O(\Upsilon^2+\kappa^2)$ iterations with high probability, with only a slightly slower rate of $(\Upsilon \log t/\sqrt{t})^t$.

One of the future works is to apply our Hessian averaging technique on constrained nonlinear optimization problems. [45, 46] designed various stochastic second-order methods based on sequential quadratic programming (SQP) for solving constrained problems, where the Hessian of the Lagrangian function was estimated by subsampling. These works established the global convergence for stochastic SQP methods, while the local convergence rate of these methods remains unknown. On the other hand, based on our analysis, the local rate of Hessian averaging schemes is induced by the central limit theorem, which we can also apply on the noise of the Lagrangian Hessian oracles. Thus, it is possible to design a stochastic SQP method with Hessian averaging and show a similar local superlinear rate for solving constrained nonlinear optimization problems. We note that a recent work [44] utilized the Hessian averaging to prove a local sublinear rate for stochastic SQP methods, while that result is not as strong as the one in this paper. Further, for unconstrained convex optimization problems, generalizing our analysis to enable stochastic gradients and function evaluations as well as inexact Newton directions is also an important future work, which can further improve the applicability of the algorithm. Finally, given any iteration threshold $\mathcal{T}$, establishing the probability that the first/second/third transition occurs before $\mathcal{T}$ is an interesting open question.

A Examples of sub-exponential noise

Lemma 8 *Consider subsampled Newton as in (6) with sample size s and suppose $\|\nabla^2 f_i(\mathbf{x})\| \leq \lambda_{\max} R$ for some $R > 0$ and for all i. Then, we have $\Upsilon = O(\kappa R\sqrt{\log(d)/s})$. If we additionally assume all f_i are convex, then we can improve this to $\Upsilon = O(\sqrt{\kappa R \log(d)/s} + \kappa R \log(d)/s)$.*

Proof The argument follows from the matrix Chernoff and Hoeffding inequalities [58, Theorems 1.1 and 1.3]. Note that the standard version of these concentration inequalities applies to sampling with replacement, namely

$$\hat{\mathbf{H}}(\mathbf{x}) = \frac{1}{s}\sum_{j=1}^{s} \nabla^2 f_{I_j}(\mathbf{x}),$$

where indices $I_1, I_2, \ldots, I_s$ are sampled from an index set uniformly at random. Alternate matrix concentration inequalities for sampling without replacement are also known, e.g., see [29, 60]. We thus take sampling with replacement as an example.

We start with the case where f_i may not be convex. By the matrix Hoeffding inequality (Theorem 1.3, [58]), we have for some small constants $c, c', c'' > 0$ and any $\eta > 0$:

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{\mathbf{H}}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \geq \eta\big) \\
&\quad \leq 2d \exp\left(-\frac{c\eta^2 s}{\lambda_{\max}^2 R^2}\right) = \exp\left(-\log(2d)\left(\frac{c\eta^2 s}{\lambda_{\max}^2 R^2 \log(2d)} - 1\right)\right) \\
&\quad \leq 2\exp\left(-\frac{c'\eta^2 s}{\lambda_{\max}^2 R^2 \log(2d)}\right) \leq 2\exp\left(-\frac{c''\eta}{\lambda_{\max} R\sqrt{\log(d)/s}}\right),
\end{aligned}$$

where the last step follows because if the expression in the exponent is larger than $-\log(2)$, then the bound is vacuous, since the probability must lie in $[0, 1]$. Thus, it follows that $\|\hat{\mathbf{H}}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| = \|\mathbf{E}(\mathbf{x})\|$ is a sub-exponential random variable with parameter $\Upsilon_E = O(\lambda_{\max} R\sqrt{\log(d)/s})$ (see Sect. 2.7 of [59]). The claim now easily follows.

Next, suppose that f_i are convex. This means that all $\nabla^2 f_i(\mathbf{x})$ are positive semidefinite, and so is $\hat{\mathbf{H}}(\mathbf{x})$. Thus we can use the matrix Chernoff inequality (Theorem 1.1 [58]), which provides a sharper guarantee:

$$\begin{aligned}
&\mathbb{P}\big(\|\hat{\mathbf{H}}(\mathbf{x}) - \mathbf{H}(\mathbf{x})\| \geq \eta\big) \\
&\quad \leq 2d \exp\left(-c \min\left\{\frac{\eta^2 s}{\lambda_{\min}\lambda_{\max} R}, \frac{\eta s}{\lambda_{\max} R}\right\}\right) \\
&\quad \leq 2\exp\left(-c' \min\left\{\frac{\eta^2 s}{\lambda_{\min}\lambda_{\max} R\log(d)}, \frac{\eta s}{\lambda_{\max} R\log(d)}\right\}\right) \\
&\quad \leq 2\exp\left(-c'' \min\left\{\frac{\eta}{\sqrt{\lambda_{\min}\lambda_{\max} R\log(d)/s}}, \frac{\eta}{\lambda_{\max} R\log(d)/s}\right\}\right)
\end{aligned}$$

$$\leq 2\exp\Big(-\frac{c''\eta}{\sqrt{\lambda_{\min}\lambda_{\max}R\log(d)/s}+\lambda_{\max}R\log(d)/s}\Big).$$

It follows that $\|\mathbf{E}(\mathbf{x})\|$ is a sub-exponential random variable with parameter $\Upsilon_E = O(\lambda_{\max}R\log(d)/s + \sqrt{\lambda_{\min}\lambda_{\max}R\log(d)/s}) = O(\lambda_{\min}(\sqrt{\kappa R\log(d)/s} + \kappa R\log(d)/s))$, and we get the claim. □

Lemma 9 *Consider Newton sketch as in (7) with* $\mathbf{S}\in\mathbb{R}^{s\times n}$ *consisting of i.i.d.* $\mathcal{N}(0,1/s)$ *entries. Then, we have* $\Upsilon = O(\kappa(\sqrt{d/s}+d/s))$.

Proof In fact, the result holds as long as $\sqrt{s}\,\mathbf{S}$ has independent mean zero unit variance sub-Gaussian entries. From a standard result on the concentration of sub-Gaussian matrices (Theorem 4.6.1 in [59]), there exists a constant $c\geq 1$ such that, with probability $1-2\exp(-t^2)$,

$$\|\mathbf{H}(\mathbf{x})^{-1/2}\hat{\mathbf{H}}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}-\mathbf{I}\|\leq\delta\vee\delta^2 \qquad \text{for any} \quad \delta\geq c\left(\sqrt{d/s}+t/\sqrt{s}\right). \tag{A.1}$$

Thus, for any $\eta>0$ such that $\eta\wedge\sqrt{\eta}\geq c\sqrt{d/s} \Longleftarrow \eta\geq c^2(d/s+\sqrt{d/s})$, we let $t=\sqrt{s}(\eta\wedge\sqrt{\eta})/c$ and have

$$c\left(\sqrt{d/s}+t/\sqrt{s}\right)=c\sqrt{d/s}+\eta\wedge\sqrt{\eta}\leq 2(\eta\wedge\sqrt{\eta}).$$

Plugging $\delta=2(\eta\wedge\sqrt{\eta})$ into (A.1), we obtain for a small constant $c'>0$ that

$$\begin{aligned}P\left(\|\mathbf{H}(\mathbf{x})^{-1/2}\hat{\mathbf{H}}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}-\mathbf{I}\|\geq 4\eta\right)&\leq 2\exp(-s\cdot(\eta\wedge\eta^2)/c^2)\\&\leq 2\exp\Big(-\frac{c'\eta}{\sqrt{1/s}+1/s}\Big).\end{aligned}$$

This further implies that, for a small constant $c''>0$,

$$\mathbb{P}(\|\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}\|\geq\eta)\leq 2\exp(-\frac{c''\eta}{\sqrt{d/s}+d/s}),\quad\forall\eta>0.$$

Thus, $\|\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}\|$ is sub-exponential with constant $K=O(\sqrt{d/s}+d/s)$. The claim follows by noting that $\|\mathbf{E}(\mathbf{x})\|\leq\lambda_{\max}\cdot\|\mathbf{H}(\mathbf{x})^{-1/2}\mathbf{E}(\mathbf{x})\mathbf{H}(\mathbf{x})^{-1/2}\|$. □

B Preliminary lemmas

Lemma 10 *Let* $d\geq 1$ *and* $\delta\in(0,1)$. *If* $d/\delta\geq e$, *then for any* $0<a\leq 1$,

$$t\geq\frac{2\log(d/(\delta a))}{a}\quad\Longrightarrow\quad\frac{\log(dt/\delta)}{t}\leq a,$$

and moreover, the right hand side inequality fails if $t=\log(d/(\delta a))/a$ *and* $d/\delta>e$.

Proof Let $t = 2\log(d/(\delta a))/a$. We will show for this t that $t \geq \log(dt/\delta)/a$. We have

$$\begin{aligned} t - \frac{\log(dt/\delta)}{a} &= \frac{2\log(d/(\delta a))}{a} - \frac{\log(d/(\delta a)) + \log(2\log(d/(\delta a)))}{a} \\ &= \frac{\log(d/(\delta a)) - \log(2\log(d/(\delta a)))}{a} \\ &= \frac{1}{a}\log\left(\frac{x}{2\log(x)}\right) \geq 0, \qquad \text{for} \quad x = \frac{d}{\delta a}, \end{aligned}$$

where the last step is due to the fact that $x \geq 2\log(x)$ for all $x > 0$. Now, we consider the function $f(t) = \log(dt/\delta)/t$. For any $t \geq 1$, its derivative is bounded as

$$\frac{\partial f(t)}{\partial t} = \frac{1 - \log(dt/\delta)}{t^2} \leq 0,$$

which implies that for $t \geq 2\log(d/(\delta a))/a \geq 1$ we get $\log(dt/\delta)/t = f(t) \leq f(2\log(d/(\delta a))/a) \leq a$. Finally, if $t = \log(d/(\delta a))/a$ and $d/\delta > e$, then $\log(dt/\delta)/a = \log(d/(\delta a))/a + \log(\log(d/(\delta a)))/a > t$, which completes the proof. □

Lemma 11 *Define the neighborhood* $\bar{\mathcal{N}}_\nu = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^\star) + \nu\lambda_{\min}^3/L^2\}$ *for* $\nu > 0$. *Suppose Assumption 2 holds, then the following relation holds*

$$\bar{\mathcal{N}}_{\nu^2/3} \subseteq \mathcal{N}_\nu \subseteq \bar{\mathcal{N}}_{\nu^2}, \quad \textit{for any } \nu \in (0, 1]$$

where $\mathcal{N}_\nu$ *is defined in* (18).

Proof By Assumption 2, we have

$$\left| f(\mathbf{x}) - f(\mathbf{x}^\star) - \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^\star\right\|_{\mathbf{H}^\star}^2 \right| \leq \frac{L}{6}\left\|\mathbf{x} - \mathbf{x}^\star\right\|^3. \tag{B.1}$$

Suppose $\mathbf{x} \in \bar{\mathcal{N}}_{\nu^2/3}$ for $\nu \in (0, 1]$, we have

$$f(\mathbf{x}^\star) + \frac{\lambda_{\min}}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2 \leq f(\mathbf{x}) \leq f(\mathbf{x}^\star) + \frac{\nu^2\lambda_{\min}^3}{3L^2} \Longrightarrow \|\mathbf{x} - \mathbf{x}^\star\| \leq \frac{\sqrt{2}\nu\lambda_{\min}}{\sqrt{3}L}. \tag{B.2}$$

Combining (B.1) and (B.2), we know that $\mathbf{x} \in \bar{\mathcal{N}}_{\nu^2/3}$ for $\nu \in (0, 1]$ leads to

$$\begin{aligned} & f(\mathbf{x}^\star) + \frac{\nu^2\lambda_{\min}^3}{3L^2} \\ & \geq f(\mathbf{x}) \overset{(B.1)}{\geq} f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{H}^\star}^2 - \frac{L}{6}\|\mathbf{x} - \mathbf{x}^\star\|^3 \\ & \geq f(\mathbf{x}^\star) + \frac{1}{2}\left\|\mathbf{x} - \mathbf{x}^\star\right\|_{\mathbf{H}^\star}^2 - \frac{L}{6}\left\|\mathbf{x} - \mathbf{x}^\star\right\| \cdot \frac{\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{H}^\star}^2}{\lambda_{\min}} \end{aligned}$$

$$= f(\mathbf{x}^\star) + \left(\frac{1}{2} - \frac{L\|\mathbf{x}-\mathbf{x}^\star\|}{6\lambda_{\min}}\right)\|\mathbf{x}-\mathbf{x}^\star\|^2_{\mathbf{H}^\star}$$
$$\overset{(B.2)}{\geq} f(\mathbf{x}^\star) + \left(\frac{1}{2} - \frac{\sqrt{2}\nu}{6\sqrt{3}}\right)\|\mathbf{x}-\mathbf{x}^\star\|^2_{\mathbf{H}^\star} \geq f(\mathbf{x}^\star) + \frac{\|\mathbf{x}-\mathbf{x}^\star\|^2_{\mathbf{H}^\star}}{3},$$

which further implies $\|\mathbf{x}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \nu\lambda_{\min}^{3/2}/L$ and $\mathbf{x} \in \mathcal{N}_\nu$. On the other hand, suppose $\mathbf{x} \in \mathcal{N}_\nu$, we have

$$f(\mathbf{x}) \overset{(B.1)}{\leq} f(\mathbf{x}^\star) + \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}^\star\right\|^2_{\mathbf{H}^\star} + \frac{L}{6}\left\|\mathbf{x}-\mathbf{x}^\star\right\|^3$$
$$\leq f(\mathbf{x}^\star) + \left(\frac{1}{2} + \frac{L\left\|\mathbf{x}-\mathbf{x}^\star\right\|_{\mathbf{H}^\star}}{6\lambda_{\min}^{3/2}}\right)\left\|\mathbf{x}-\mathbf{x}^\star\right\|^2_{\mathbf{H}^\star}$$
$$\leq f(\mathbf{x}^\star) + \left(\frac{1}{2} + \frac{\nu}{6}\right)\left\|\mathbf{x}-\mathbf{x}^\star\right\|^2_{\mathbf{H}^\star} \leq f(\mathbf{x}^\star) + \left\|\mathbf{x}-\mathbf{x}^\star\right\|^2_{\mathbf{H}^\star} \leq f(\mathbf{x}^\star) + \frac{\nu^2\lambda_{\min}^3}{L^2}.$$

Thus, we have $\mathbf{x} \in \bar{\mathcal{N}}_{\nu^2}$. This completes the proof. □

Lemma 12 *Suppose Assumption 2 holds, then we have*

$$\|\mathbf{H}(\mathbf{x}_1)^{-1/2}\mathbf{H}(\mathbf{x}_2)\mathbf{H}(\mathbf{x}_1)^{-1/2} - \mathbf{I}\| \leq \frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_1-\mathbf{x}_2\|_{\mathbf{H}^\star}, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d.$$

Proof We note that

$$\|\mathbf{H}(\mathbf{x}_1)^{-1/2}\mathbf{H}(\mathbf{x}_2)\mathbf{H}(\mathbf{x}_1)^{-1/2} - \mathbf{I}\| \leq \frac{1}{\lambda_{\min}}\|\mathbf{H}(\mathbf{x}_1) - \mathbf{H}(\mathbf{x}_2)\|$$
$$\leq \frac{L}{\lambda_{\min}}\|\mathbf{x}_1-\mathbf{x}_2\| \leq \frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_1-\mathbf{x}_2\|_{\mathbf{H}^\star}.$$

□

C Proofs

C.1 Proof of Theorem 3

For any $t \geq 0$ and scalar $\theta > 0$, we have from Assumption 1 that

$$\mathbb{E}_i\left[\exp\left(\theta z_{i,t}\mathbf{E}_i\right)\right]$$
$$= \mathbf{I} + \theta z_{i,t}\mathbb{E}_i[\mathbf{E}_i] + \sum_{j=2}^{\infty}\frac{(\theta z_{i,t})^j\mathbb{E}_i[\mathbf{E}_i^j]}{j!}$$
$$\preceq \mathbf{I} + \sum_{j=2}^{\infty}(\theta z_{i,t}\Upsilon_E)^{j-2}\cdot\frac{(\theta z_{i,t}\Upsilon_E)^2}{2}\cdot\mathbf{I}$$

$$= \mathbf{I} + \frac{(\theta z_{i,t}\Upsilon_E)^2}{2(1-\theta z_{i,t}\Upsilon_E)} \cdot \mathbf{I} \preceq \exp\left(\frac{(\theta z_{i,t}\Upsilon_E)^2}{2(1-\theta z_{i,t}\Upsilon_E)} \cdot \mathbf{I}\right), \quad \forall i = 0, 1, \ldots, t,$$

where the second equality holds for $\left\{\theta : \theta z_{i,t}\Upsilon_E < 1\right\}$. Therefore, we let $\Theta_t := \{\theta : \theta < 1/(z_t^{(\max)}\Upsilon_E)\}$, and have for all $i = 0, 1, \ldots, t$,

$$\mathbb{E}_i\left[\exp\left(\theta z_{i,t}\mathbf{E}_i\right)\right] \preceq \exp\left(\frac{(\theta z_{i,t}\Upsilon_E)^2}{2(1-\theta z_t^{(\max)}\Upsilon_E)} \cdot \mathbf{I}\right), \quad \forall \theta \in \Theta_t.$$

The above inequality suggests that the condition (12) in Lemma 1 holds with

$$g_t(\theta) = \frac{\theta^2}{2(1-\theta z_t^{(\max)}\Upsilon_E)} \quad \text{and} \quad \mathbf{U}_i = z_{i,t}^2\Upsilon_E^2 \cdot \mathbf{I}.$$

Let $\sigma^2 = \|\sum_{i=0}^t \mathbf{U}_i\| = \Upsilon_E^2 \sum_{i=0}^t z_{i,t}^2$ in Lemma 1. Then we have for any $\eta \geq 0$,

$$\mathbb{P}\left(\left\|\bar{\mathbf{E}}_t\right\| \geq \eta\right) \leq 2d \cdot \inf_{\theta \in \Theta_t} \exp\left(-\theta\eta + g_t(\theta)\sigma^2\right).$$

Plugging $\theta = \eta/(z_t^{(\max)}\Upsilon_E\eta + \sigma^2) \in \Theta_t$ into the above right hand side, we obtain

$$\mathbb{P}\left(\|\bar{\mathbf{E}}_t\| \geq \eta\right) \leq 2d \cdot \exp\left(-\frac{\eta^2/2}{\sigma^2 + z_t^{(\max)}\Upsilon_E\eta}\right).$$

This completes the proof.

C.2 Proof of Lemma 2

Let us define the event

$$\bar{\mathcal{E}} = \bigcap_{t=0}^{\infty}\left\{\left\|\bar{\mathbf{E}}_t\right\| \leq 8\Upsilon_E\left(\sqrt{\frac{\log(d(t+1)/\delta)}{t+1}} \vee \frac{\log(d(t+1)/\delta)}{t+1}\right)\right\}.$$

By Theorem 3 and (15), we know $\mathbb{P}(\bar{\mathcal{E}}) \geq 1 - \delta\pi^2/6$. We also note that

$$\begin{aligned}
\|\bar{\mathbf{E}}_t\| \leq \epsilon\lambda_{\min} &\Longleftarrow 8\Upsilon_E\left(\sqrt{\frac{\log(d(t+1)/\delta)}{t+1}} \vee \frac{\log(d(t+1)/\delta)}{t+1}\right) \leq \epsilon\lambda_{\min} \\
&\Longleftarrow \frac{\log(d(t+1)/\delta)}{t+1} \leq \frac{\epsilon^2}{64\Upsilon^2} \wedge 1 = \left(\frac{\epsilon}{8\Upsilon} \wedge 1\right)^2 \\
&\Longleftarrow t \geq 4\left(\frac{8\Upsilon}{\epsilon} \vee 1\right)^2 \log\left\{\frac{d}{\delta} \cdot \left(\frac{8\Upsilon}{\epsilon} \vee 1\right)\right\} = \mathcal{T}_1. \quad \text{(Lemma 10)}
\end{aligned} \tag{C.1}$$

Therefore, with probability at least $1 - \delta\pi^2/6$, we have for any $t \geq \mathcal{T}_1$ that $\|\bar{\mathbf{E}}_t\| \leq \epsilon\lambda_{\min}$. This shows that the event $\mathcal{E}$ defined in (17) happens, and

$$(1-\epsilon)\lambda_{\min} \cdot \mathbf{I} \preceq \tilde{\mathbf{H}}_t \overset{(11)}{=} \frac{1}{t+1}\sum_{i=0}^{t} \mathbf{H}_i + \bar{\mathbf{E}}_t \preceq (\lambda_{\max} + \epsilon\lambda_{\min}) \cdot \mathbf{I} \preceq (1+\epsilon)\lambda_{\max} \cdot \mathbf{I}.$$

This completes the proof.

C.3 Proof of Lemma 3

By Lemma 2, we know that $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$ for $t \geq \mathcal{T}_1$. We apply Taylor's expansion:

$$\begin{aligned}
&f(\mathbf{x}_t - \mu_t(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t) \\
&\leq f(\mathbf{x}_t) - \mu_t \nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{\lambda_{\max}\mu_t^2}{2}\nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-2}\nabla f_t \\
&\leq f(\mathbf{x}_t) - \mu_t \nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{\kappa\mu_t^2}{2(1-\epsilon)}\nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \quad \text{(Lemma 2)} \\
&= f(\mathbf{x}_t) - \mu_t\left(1 - \frac{\kappa\mu_t}{2(1-\epsilon)}\right)\nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t.
\end{aligned}$$

Thus, the Armijo condition is satisfied if

$$1 - \frac{\kappa\mu_t}{2(1-\epsilon)} \geq \beta \iff \mu_t \leq \frac{2(1-\beta)(1-\epsilon)}{\kappa}.$$

Therefore, the backtracking line search on Line 7 leads to a stepsize satisfying

$$\mu_t \geq \frac{2\rho(1-\beta)(1-\epsilon)}{\kappa}. \tag{C.2}$$

Moreover, we apply Lemma 2 and strong convexity of f, and have

$$\nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \geq \frac{1}{(1+\epsilon)\lambda_{\max}}\|\nabla f_t\|^2 \geq \frac{2}{(1+\epsilon)\kappa}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)). \tag{C.3}$$

Combining (C.2), (C.3) with the Armijo condition, we have

$$\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) &\leq f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \mu_t\beta\nabla f_t^\top (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \\
&\leq f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \frac{2\rho\beta(1-\beta)(1-\epsilon)}{\kappa} \cdot \frac{2}{(1+\epsilon)\kappa}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \\
&= (1-\phi)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)),
\end{aligned} \tag{C.4}$$

which shows the first part of the statement. Furthermore, we apply (C.4) recursively, apply strong convexity, and have

$$\begin{aligned}\|\mathbf{x}_t-\mathbf{x}^\star\|^2 &\leq \frac{2(f(\mathbf{x}_t)-f(\mathbf{x}^\star))}{\lambda_{\min}}\\ &\overset{(C.4)}{\leq} \frac{2(1-\phi)^{t-\mathcal{T}_1}}{\lambda_{\min}}(f(\mathbf{x}_{\mathcal{T}_1})-f(\mathbf{x}^\star))\\ &\leq \frac{2(1-\phi)^{t-\mathcal{T}_1}}{\lambda_{\min}}(f(\mathbf{x}_0)-f(\mathbf{x}^\star)).\end{aligned}$$

The last statement follows from $\|\mathbf{x}_t-\mathbf{x}^\star\|^2_{\mathbf{H}^\star}\leq\lambda_{\max}\|\mathbf{x}_t-\mathbf{x}^\star\|^2$. This completes the proof.

C.4 Proof of Corollary 1

We condition on the event (17), which happens with probability $1-\delta\pi^2/6$. By Lemma 3, we know that, for $t\geq\mathcal{T}_1$,

$$f(\mathbf{x}_t)-f(\mathbf{x}^\star)\leq(1-\phi)^{t-\mathcal{T}_1}(f(\mathbf{x}_{\mathcal{T}_1})-f(\mathbf{x}^\star))\leq(1-\phi)^{t-\mathcal{T}_1}(f(\mathbf{x}_0)-f(\mathbf{x}^\star)).$$

By Lemma 11, to have $\mathbf{x}_t\in\mathcal{N}_\nu$, it suffices to have $\mathbf{x}_t\in\bar{\mathcal{N}}_{\nu^2/3}$. Thus, we let

$$\begin{aligned}&(1-\phi)^{t-\mathcal{T}_1}(f(\mathbf{x}_0)-f(\mathbf{x}^\star))\leq\frac{\nu^2\lambda_{\min}^3}{3L^2}\\ \Longleftarrow\ & t-\mathcal{T}_1\geq\frac{\log\left(\frac{3L^2(f(\mathbf{x}_0)-f(\mathbf{x}^\star))}{\nu^2\lambda_{\min}^3}\right)}{\log\left(\frac{1}{1-\phi}\right)}\Longleftarrow t\geq\mathcal{T}_1+\mathcal{T}_2,\end{aligned}\tag{C.5}$$

where $\mathcal{T}_2$ is defined in (19) and the implication uses the fact that $\log(1/(1-\phi))\geq\phi$. Furthermore, since $f(\mathbf{x}_t)$ is always decreasing, we know after $\mathcal{T}_1+T_2$ iterations $\mathbf{x}_t$ stays in $\bar{\mathcal{N}}_{\nu^2/3}$, and hence stays in $\mathcal{N}_\nu$ (cf. Lemma 11). This completes the proof.

C.5 Proof of Lemma 4

It suffices to show

$$f(\mathbf{x}_t-(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t)\leq f(\mathbf{x}_t)-\beta\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t.$$

By Assumption 2, we have

$$\begin{aligned}&f(\mathbf{x}_t-(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t)\\ &\leq f(\mathbf{x}_t)-\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t+\frac{1}{2}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\end{aligned}$$

$$
\begin{aligned}
&+ \frac{L}{6}\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|^3 \\
&\le f(\mathbf{x}_t) - \frac{1}{2}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{1}{2}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}(\mathbf{H}_t - \tilde{\mathbf{H}}_t)(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{L}{6}\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|^3 \\
&\le f(\mathbf{x}_t) - \frac{1}{2}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{1}{2}\|(\tilde{\mathbf{H}}_t)^{-1/2}(\mathbf{H}_t - \tilde{\mathbf{H}}_t)(\tilde{\mathbf{H}}_t)^{-1/2}\|\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \\
&+ \frac{L}{6}\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-2}\nabla f_t \\
&\le f(\mathbf{x}_t) - \frac{1}{2}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t + \frac{\psi}{2(1-\psi)}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \\
&+ \frac{L\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|}{6(1-\psi)\lambda_{\min}}\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t \\
&= f(\mathbf{x}_t) - \left(\frac{1}{2} - \frac{\psi}{2(1-\psi)} - \frac{L\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|}{6(1-\psi)\lambda_{\min}}\right)\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t. \qquad \text{(C.6)}
\end{aligned}
$$

For term $\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|$, we let $\mathbf{H}_t^\eta = \mathbf{H}(\mathbf{x}_t^\eta)$ with $\mathbf{x}_t^\eta = \mathbf{x}^\star + \eta(\mathbf{x}_t - \mathbf{x}^\star)$ for some $\eta \in (0, 1)$, and have

$$
\begin{aligned}
\|(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|^2 &= (\mathbf{x}_t - \mathbf{x}^\star)^\top(\mathbf{H}^\star)^{1/2}(\mathbf{H}^\star)^{-1/2}\mathbf{H}_t^\eta(\tilde{\mathbf{H}}_t)^{-2}\mathbf{H}_t^\eta(\mathbf{H}^\star)^{-1/2}(\mathbf{H}^\star)^{1/2}(\mathbf{x}_t - \mathbf{x}^\star) \\
&\le \frac{\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}^2}{(1-\psi)\lambda_{\min}}\|(\mathbf{H}^\star)^{-1/2}\mathbf{H}_t^\eta(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^\eta(\mathbf{H}^\star)^{-1/2}\| \\
&\le \frac{\nu^2\lambda_{\min}^2}{(1-\psi)L^2}\|(\mathbf{H}^\star)^{-1/2}(\mathbf{H}_t^\eta)^{1/2}\|^2\|(\mathbf{H}_t^\eta)^{1/2}(\tilde{\mathbf{H}}_t)^{-1}(\mathbf{H}_t^\eta)^{1/2}\| \quad (\text{since } \mathbf{x}_t \in \mathcal{N}_\nu) \\
&\le \frac{\nu^2\lambda_{\min}^2}{(1-\psi)L^2}\|(\mathbf{H}^\star)^{-1/2}\mathbf{H}_t^\eta(\mathbf{H}^\star)^{-1/2}\|\|(\mathbf{H}_t^\eta)^{1/2}(\mathbf{H}_t)^{-1}(\mathbf{H}_t^\eta)^{1/2}\|\|\mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\| \\
&\le \frac{\nu^2\lambda_{\min}^2}{(1-\psi)^2L^2}\|(\mathbf{H}^\star)^{-1/2}\mathbf{H}_t^\eta(\mathbf{H}^\star)^{-1/2}\|\|(\mathbf{H}_t)^{-1/2}\mathbf{H}_t^\eta(\mathbf{H}_t)^{-1/2}\|. \qquad \text{(C.7)}
\end{aligned}
$$

Noting that $\mathbf{x}_t^\eta \in \mathcal{N}_\nu$ if $\mathbf{x}_t \in \mathcal{N}_\nu$, we apply Lemma 12 and have

$$
\|(\mathbf{H}^\star)^{-1/2}\mathbf{H}_t^\eta(\mathbf{H}^\star)^{-1/2}\| \vee \|(\mathbf{H}_t)^{-1/2}\mathbf{H}_t^\eta(\mathbf{H}_t)^{-1/2}\| \le 1 + \nu \le 2. \qquad \text{(C.8)}
$$

Combining (C.6), (C.7), and (C.8), we finally obtain

$$
f(\mathbf{x}_t - (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t) \le f(\mathbf{x}_t) - \left(\frac{1}{2} - \frac{\psi}{2(1-\psi)} - \frac{\nu}{3(1-\psi)^2}\right)\nabla f_t^\top(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t.
$$

Thus, it suffices to let

$$
\frac{1}{2} - \frac{\psi}{2(1-\psi)} - \frac{\nu}{3(1-\psi)^2} \ge \beta \Longleftarrow \frac{\psi}{2(1-\psi)} \vee \frac{\nu}{3(1-\psi)^2} \le \frac{1}{2}\left(\frac{1}{2} - \beta\right)
$$

as implied by the conditions stated in the lemma. This completes the proof.

C.6 Proof of Lemma 5

Since $\mathbf{x}_t \in \mathcal{N}_\nu$ with $\nu \leq 2/3 \cdot (0.5 - \beta)$, we apply Lemma 12 and have

$$\|(\mathbf{H}^\star)^{-1/2}(\mathbf{H}_t - \mathbf{H}^\star)(\mathbf{H}^\star)^{-1/2}\| \leq \frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \nu \leq 0.5 - \beta.$$

Thus, we obtain

$$(0.5 + \beta)\,\mathbf{H}^\star \preceq \mathbf{H}_t \preceq (1.5 - \beta)\,\mathbf{H}^\star. \tag{C.9}$$

Since $\mathbf{p}_t = -(\tilde{\mathbf{H}}_t)^{-1}\nabla f_t$ and $\mu_t = 1$, we have

$$\begin{aligned}
\left\|\mathbf{x}_{t+1} - \mathbf{x}^\star\right\|_{\mathbf{H}^\star} &= \|\mathbf{x}_t - (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}^\star} \\
&\overset{(C.9)}{\leq} \frac{1}{\sqrt{0.5+\beta}}\|\mathbf{x}_t - (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}_t} \\
&\leq \frac{1}{\sqrt{0.5+\beta}}\left\{\|\mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}_t} + \|\mathbf{H}_t^{-1}\nabla f_t - (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|_{\mathbf{H}_t}\right\}. \qquad \text{(C.10)}
\end{aligned}$$

For the second term on the right hand side, we have

$$\begin{aligned}
&\|\mathbf{H}_t^{-1}\nabla f_t - (\tilde{\mathbf{H}}_t)^{-1}\nabla f_t\|_{\mathbf{H}_t} \\
&= \sqrt{\nabla f_t^\top(\mathbf{H}_t^{-1} - (\tilde{\mathbf{H}}_t)^{-1})\mathbf{H}_t(\mathbf{H}_t^{-1} - (\tilde{\mathbf{H}}_t)^{-1})\nabla f_t} \\
&= \sqrt{\nabla f_t^\top \mathbf{H}_t^{-1/2}(\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2})^2\mathbf{H}_t^{-1/2}\nabla f_t} \\
&\leq \|\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\| \cdot \|\mathbf{H}_t^{-1}\nabla f_t\|_{\mathbf{H}_t} \\
&\leq \|\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\|\left\{\|\mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}_t} + \|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}_t}\right\} \\
&\overset{(C.9)}{\leq} \|\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\|\left\{\|\mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}_t} + \sqrt{1.5-\beta}\|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}\right\}. \qquad \text{(C.11)}
\end{aligned}$$

Combining (C.10) and (C.11),

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \frac{1}{\sqrt{0.5+\beta}}\big\{&(1 + \|\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\|)\|\mathbf{x}_t - \mathbf{H}_t^{-1}\nabla f_t - \mathbf{x}^\star\|_{\mathbf{H}_t} \\
&+ \sqrt{1.5-\beta}\|\mathbf{I} - \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\| \cdot \|\mathbf{x}_t - \mathbf{x}^\star\|_{\mathbf{H}^\star}\big\}. \qquad \text{(C.12)}
\end{aligned}$$

Since $\|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\| \leq \psi \leq 1/3$, we have

$$(1 - \|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|) \cdot \mathbf{I} \preceq \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2} \preceq (1 + \|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|) \cdot \mathbf{I}$$

and further

$$-\frac{\|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|}{1 + \|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|} \cdot \mathbf{I} \preceq \mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2} - \mathbf{I} \preceq \frac{\|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|}{1 - \|\mathbf{I} - \mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|} \cdot \mathbf{I}.$$

Thus, we have

$$\|\mathbf{I}-\mathbf{H}_t^{1/2}(\tilde{\mathbf{H}}_t)^{-1}\mathbf{H}_t^{1/2}\| \le \frac{\|\mathbf{I}-\mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|}{1-\|\mathbf{I}-\mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|} \le \frac{3}{2}\|\mathbf{I}-\mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|.$$

Combining the above inequality with (C.12), we obtain

$$\begin{aligned}
&\|\mathbf{x}_{t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\\
&\le \frac{3}{2\sqrt{0.5+\beta}}\|\mathbf{x}_t-\mathbf{H}_t^{-1}\nabla f_t-\mathbf{x}^\star\|_{\mathbf{H}_t}\\
&+\frac{3}{2}\sqrt{\frac{1.5-\beta}{0.5+\beta}}\cdot\|\mathbf{I}-\mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|\cdot\|\mathbf{x}_t-\mathbf{x}^\star\|_{\mathbf{H}^\star}\\
&\le \frac{3\sqrt{2}}{2}\|\mathbf{x}_t-\mathbf{H}_t^{-1}\nabla f_t-\mathbf{x}^\star\|_{\mathbf{H}_t}+\frac{3\sqrt{3}}{2}\|\mathbf{I}-\mathbf{H}_t^{-1/2}\tilde{\mathbf{H}}_t\mathbf{H}_t^{-1/2}\|\cdot\|\mathbf{x}_t-\mathbf{x}^\star\|_{\mathbf{H}^\star}.
\end{aligned} \tag{C.13}$$

Furthermore,

$$\begin{aligned}
&\|\mathbf{x}_t-\mathbf{H}_t^{-1}\nabla f_t-\mathbf{x}^\star\|_{\mathbf{H}_t}\\
&=\|\mathbf{H}_t^{-1/2}(\mathbf{H}_t(\mathbf{x}_t-\mathbf{x}^\star)-\nabla f_t)\|\\
&\le \frac{1}{\sqrt{\lambda_{\min}}}\left\|\mathbf{H}_t(\mathbf{x}_t-\mathbf{x}^\star)-\left(\int_0^1\mathbf{H}(\mathbf{x}^\star+\tau(\mathbf{x}_t-\mathbf{x}^\star))d\tau\right)(\mathbf{x}_t-\mathbf{x}^\star)\right\|\\
&\le \frac{\|\mathbf{x}_t-\mathbf{x}^\star\|^2}{\sqrt{\lambda_{\min}}}\cdot\int_0^1(1-\tau)Ld\tau \le \frac{L}{2\lambda_{\min}^{3/2}}\|\mathbf{x}_t-\mathbf{x}^\star\|_{\mathbf{H}^\star}^2.
\end{aligned}$$

Combining the above inequality with (C.13), we complete the proof.

C.7 Proof of Theorem 4

We suppose the event (17) happens, which has probability $1-\delta\pi^2/6$. By Lemma 3 and Corollary 1, we know that $\mathbf{x}_t$ converges R-linearly for all $t \ge \mathcal{T}_1$, and $\mathbf{x}_{\mathcal{T}:\mathcal{T}+\mathcal{J}} \in \mathcal{N}_v$ for any $\mathcal{J} \ge 0$. Thus, it suffices to study the convergence after $\mathcal{T}+\mathcal{J}$. For any $t \ge 0$, to apply Lemmas 4 and 5, we characterize the difference between $\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}$ and $\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}$. We have

$$\begin{aligned}
&\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}-\mathbf{I}\|\\
&\le \|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\|+\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\mathbf{H}^\star\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}-\mathbf{I}\|\\
&\overset{\text{Lemma 12}}{\le} \|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\|+\frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}.
\end{aligned} \tag{C.14}$$

For the first term $\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\|$, we apply the formula (11) and have

$$
\begin{aligned}
&\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\| \\
&\overset{(11)}{\leq} \frac{1}{\lambda_{\min}}\left\|\bar{\mathbf{E}}_{\mathcal{T}+\mathcal{J}+t}\right\| + \frac{1}{\mathcal{T}+\mathcal{J}+t+1}\left\{\sum_{j=0}^{\mathcal{T}-1}\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\mathbf{H}_j-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\|\right. \\
&\quad \left. + \sum_{j=\mathcal{T}}^{\mathcal{T}+\mathcal{J}+t}\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}(\mathbf{H}_j-\mathbf{H}^\star)\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\|\right\} \\
&\leq \frac{1}{\lambda_{\min}}\|\bar{\mathbf{E}}_{\mathcal{T}+\mathcal{J}+t}\| + \frac{2\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} + \frac{L}{\lambda_{\min}(\mathcal{T}+\mathcal{J}+t+1)}\sum_{j=\mathcal{T}}^{\mathcal{T}+\mathcal{J}+t}\|\mathbf{x}_j-\mathbf{x}^\star\| \\
&\overset{(17)}{\leq} 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} + \frac{2\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} \\
&\quad + \frac{L}{\lambda_{\min}(\mathcal{T}+\mathcal{J}+t+1)}\sum_{j=\mathcal{T}}^{\mathcal{T}+\mathcal{J}+t} \\
&\quad \times \sqrt{\frac{2(f(\mathbf{x}_0)-f(\mathbf{x}^\star))}{\lambda_{\min}}}(1-\phi)^{(j-\mathcal{T}_1)/2} \quad \text{(also use Lemma 3)} \\
&\leq 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} + \frac{2\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} \\
&\quad + \frac{L\sqrt{2(f(\mathbf{x}_0)-f(\mathbf{x}^\star))}}{\lambda_{\min}^{3/2}(\mathcal{T}+\mathcal{J}+t+1)}(1-\phi)^{\mathcal{T}_2/2}\sum_{j=0}^{\infty}(1-\phi)^{j/2} \\
&\overset{(C.5)}{\leq} 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} + \frac{2\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} \\
&\quad + \frac{\sqrt{2}\nu}{\sqrt{3}(\mathcal{T}+\mathcal{J}+t+1)(1-\sqrt{1-\phi})} \\
&\leq 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} \\
&\quad + \frac{2\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} + \frac{2\nu}{(\mathcal{T}+\mathcal{J}+t+1)\phi} \\
&\leq 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} \\
&\quad + \frac{4\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} = \rho_t. \quad (\text{since } \nu/\phi \leq 1/\phi \leq \mathcal{T}\kappa)
\end{aligned}
\tag{C.15}
$$

Combining (C.14) and (C.15) together, we obtain

$$\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}-\mathbf{I}\| \le \rho_t+\frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \le \rho_t+\nu \le \rho_0+\nu. \tag{C.16}$$

Thus, in order to apply Lemma 5 for all $\{\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}\}_{t\ge 0}$, we require $\nu \le 2/3\cdot(0.5-\beta)$ and

$$\begin{aligned}
\rho_0+\nu \le \frac{0.5-\beta}{1.5-\beta} &\overset{\mathcal{J}=4\mathcal{T}\kappa/\nu}{\Longleftarrow} 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J})/\delta)}{\mathcal{T}+\mathcal{J}}}+2\nu \le \frac{0.5-\beta}{1.5-\beta}\\
&\Longleftarrow 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J})/\delta)}{\mathcal{T}+\mathcal{J}}} \vee \nu \le \frac{1}{3}\frac{0.5-\beta}{1.5-\beta}\\
&\overset{(21)}{\Longleftarrow} 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J})/\delta)}{\mathcal{T}+\mathcal{J}}} \le \epsilon \text{ and } \epsilon\vee\nu \le \frac{1}{3}\frac{0.5-\beta}{1.5-\beta},
\end{aligned}$$

which is implied by (21) and (C.1). Thus, for any $t \ge 0$, we apply Lemma 5 and obtain

$$\begin{aligned}
&\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\\
&\le 3\left\{\frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}^2+\|\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{T}+\mathcal{J}+t}\mathbf{H}_{\mathcal{T}+\mathcal{J}+t}^{-1/2}-\mathbf{I}\|\cdot\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\right\}\\
&\overset{(C.16)}{\le} 3\left\{\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}^2+\rho_t\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\right\}.
\end{aligned} \tag{C.17}$$

We then claim that

$$\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \le 3\rho_t, \quad \forall t\ge 0. \tag{C.18}$$

We prove (C.18) by induction. For $t=0$, we note that

$$\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \le 2\nu = \frac{12\mathcal{T}\kappa}{6\mathcal{T}\kappa/\nu} \le \frac{12\mathcal{T}\kappa}{\mathcal{T}+4\mathcal{T}\kappa/\nu+1} = \frac{12\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+1} \le 3\rho_0.$$

Suppose (C.18) holds for $t\ge 0$, then (C.17) leads to

$$\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \le \frac{2L}{\lambda_{\min}^{3/2}}\cdot 12\rho_t\|\mathbf{x}_{\mathcal{T}+\mathcal{J}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \overset{(C.18)}{\le} 36\rho_t^2.$$

On the other hand, using $(\mathcal{T}+\mathcal{J}+t+2) \le 2(\mathcal{T}+\mathcal{J}+t+1)$, $\forall t\ge 0$, we know $\rho_t/2 \le \rho_{t+1}$. Thus, it suffices to show $36\rho_t^2 \le 3\rho_t/2 \iff 24\rho_t \le 1$. Since ρ_t is decreasing, we require $24\rho_0 \le 1$. Note that

$$\rho_0 \le 8\Upsilon\sqrt{\frac{\log(d\mathcal{T}/\delta)}{\mathcal{T}}} + \nu \overset{(19)}{\le} \epsilon + \nu.$$

Thus, $24\rho_0 \le 1$ is guaranteed by (21). Combining (C.17) and (C.18) completes the proof.

C.8 Proof of Corollary 3

We only need to note that

$$\begin{aligned}
&\frac{4\mathcal{T}\kappa}{\mathcal{T}+\mathcal{J}+t+1} \le 8\Upsilon\sqrt{\frac{\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta)}{\mathcal{T}+\mathcal{J}+t+1}} \\
&\iff \frac{\mathcal{T}^2\kappa^2}{4\Upsilon^2} \le (\mathcal{T}+\mathcal{J}+t+1)\log(d(\mathcal{T}+\mathcal{J}+t+1)/\delta) \\
&\impliedby \mathcal{T}+\mathcal{J}+t+1 \ge \frac{\mathcal{T}^2\kappa^2}{4\Upsilon^2\log(d\mathcal{T}/\delta)}.
\end{aligned}$$

This completes the proof.

C.9 Proof of Lemma 6

We characterize $\sum_{i=0}^{t} z_{i,t}^2$ and $z_t^{(\max)}$ in (14), where $z_{i,t} = (w_i - w_{i-1})/w_t$ and $z_t^{(\max)} = \max_{i\in\{0,\dots,t\}} z_{i,t}$. We have

$$\begin{aligned}
\sum_{i=0}^{t} z_{i,t}^2 &= \frac{1}{w(t)^2}\sum_{i=0}^{t}(w(i)-w(i-1))^2 \\
&\le \frac{1}{w(t)^2}\sum_{i=0}^{t}(w'(i))^2 \le \frac{1}{w(t)^2}\int_0^{t+1}(w'(i))^2\,di \\
&= \frac{1}{w(t)^2}\int_0^{t+1}(w(i)w'(i))' - w(i)w''(i)di \\
&\le \frac{w(t+1)w'(t+1)}{w(t)^2} \le \Psi^2\frac{w'(t)}{w(t)},
\end{aligned} \tag{C.19}$$

where the first two inequalities use the fact that $w'(t)$ is non-negative and non-decreasing; the second last inequality uses the fact that $w(t)w''(t) \ge 0$, $\forall t \ge -1$; and the last inequality uses Assumption 3(v). By the same derivation, we have

$$\begin{aligned}
z_t^{(\max)} &= \max_{i\in\{0,\dots,t\}} z_{i,t} = \frac{\max_{i\in\{0,\dots,t\}}(w(i)-w(i-1))}{w(t)} \\
&\le \frac{\max_{i\in\{0,\dots,t\}} w'(i)}{w(t)} = \frac{w'(t)}{w(t)}.
\end{aligned} \tag{C.20}$$

Plugging (C.19) and (C.20) into (14) completes the proof.

C.10 Proof of Lemma 7

By Lemma 6, we know (22) holds for any t. By the definition of $\mathcal{I}_1$ in (23), we know

$$\log\left(\frac{d(t+1)}{\delta}\right)\frac{w'(t)}{w(t)} \le \left(\frac{\epsilon}{8\Upsilon\Psi}\right)^2 \wedge 1, \quad \forall t \ge \mathcal{I}_1,$$

which implies $\|\bar{\mathbf{E}}_t\| \le \epsilon\lambda_{\min}$. Using (11) and noting that $\lambda_{\min}\cdot\mathbf{I} \preceq \sum_{i=0}^{t}(w_i - w_{i-1})\mathbf{H}_i/w_t \preceq \lambda_{\max}\cdot\mathbf{I}$, we complete the proof.

C.11 Proof of Theorem 5

We follow the same proof structure as Theorem 4. We suppose the event (24) happens, which occurs with probability $1-\delta\pi^2/6$. We only need to study the convergence after $\mathcal{I}+\mathcal{U}$ where $\mathcal{U}$ is chosen such that $w(\mathcal{I}+\mathcal{U}) = 2w(\mathcal{I}-1)\kappa/\nu$. We claim that

$$\begin{aligned}&\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \le 6\theta_t\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \text{ and}\\ &\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \le \theta_{t+1}, \ \forall t \ge 0.\end{aligned} \tag{C.21}$$

We prove (C.21) by induction. Before showing (C.21), we note that

$$\theta_t \le \epsilon + 3\nu \le 1/(12\Psi), \qquad \theta_{t+1} \ge \theta_t/\Psi, \qquad \forall t \ge 0. \tag{C.22}$$

The first result is implied by (26) and the second result is implied by Assumption 3(v). For $t=0$, we characterize the difference between $\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}}$ and $\mathbf{H}_{\mathcal{I}+\mathcal{U}}$. We have

$$\begin{aligned}&\|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}}\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2} - \mathbf{I}\|\\ &\quad\le \|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}} - \mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\| + \|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\mathbf{H}^\star\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2} - \mathbf{I}\|\\ &\quad\le \|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}} - \mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\| + \frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{I}+\mathcal{U}} - \mathbf{x}^\star\|_{\mathbf{H}^\star}. \quad \text{(Lemma 12)}\end{aligned} \tag{C.23}$$

For the first term on the right hand side, we have

$$\begin{aligned}&\|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}} - \mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\|\\ &\overset{(11)}{\le} \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{1}{w(\mathcal{I}+\mathcal{U})}\sum_{j=0}^{\mathcal{I}+\mathcal{U}}(w(j)-w(j-1))\|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}(\mathbf{H}_j - \mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\|\\ &= \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{1}{w(\mathcal{I}+\mathcal{U})}\end{aligned}$$

$$\times \left\{ \left(\sum_{j=0}^{\mathcal{I}-1} + \sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}} \right) (w(j) - w(j-1)) \|\mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2} (\mathbf{H}_j - \mathbf{H}^\star) \mathbf{H}_{\mathcal{I}+\mathcal{U}}^{-1/2}\| \right\}$$

$$\leq \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{2w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})} + \frac{L}{\lambda_{\min} w(\mathcal{I}+\mathcal{U})} \sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}} (w(j) - w(j-1)) \|\mathbf{x}_j - \mathbf{x}^\star\|$$

$$\leq \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{2w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})}$$

$$+ \sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}} \frac{L(w(j) - w(j-1))}{\lambda_{\min} w(\mathcal{I}+\mathcal{U})} \left\{ \frac{2(f(\mathbf{x}_0) - f(\mathbf{x}^\star))}{\lambda_{\min}} (1-\phi)^{j-\mathcal{I}_1} \right\}^{1/2} \quad \text{(Lemma 3)}$$

$$= \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{2w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})}$$

$$+ \left\{ \frac{2L^2(f(\mathbf{x}_0) - f(\mathbf{x}^\star))}{\lambda_{\min}^3} (1-\phi)^{\mathcal{I}_2} \right\}^{1/2} \sum_{j=0}^{\mathcal{U}} \frac{w(\mathcal{I}+j) - w(\mathcal{I}+j-1)}{w(\mathcal{I}+\mathcal{U})} (1-\phi)^{j/2}$$

$$\overset{(C.5)}{\leq} \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{2w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})} + \nu \sum_{j=0}^{\mathcal{U}} \frac{w(\mathcal{I}+j) - w(\mathcal{I}+j-1)}{w(\mathcal{I}+\mathcal{U})}$$

$$\leq \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{2w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})} + \nu = \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}}\|}{\lambda_{\min}} + \frac{4w(\mathcal{I}-1)\kappa}{w(\mathcal{I}+\mathcal{U})} \leq \theta_0. \tag{C.24}$$

Combining (C.23) and (C.24), we know that, to apply Lemma 5, we need $\nu \leq 2/3 \cdot (0.5 - \beta)$ and

$$\theta_0 + \nu \leq \frac{0.5-\beta}{1.5-\beta} \overset{(C.22)}{\Longleftarrow} \epsilon + 4\nu \leq \frac{0.5-\beta}{1.5-\beta} \Longleftarrow \epsilon \vee \nu \leq \frac{1}{5} \frac{0.5-\beta}{1.5-\beta},$$

as implied by (26). Thus, Lemma 5 leads to

$$\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 3 \left\{ \frac{2L}{\lambda_{\min}^{3/2}} \|\mathbf{x}_{\mathcal{I}+\mathcal{U}} - \mathbf{x}^\star\|_{\mathbf{H}^\star}^2 + \theta_0 \|\mathbf{x}_{\mathcal{I}+\mathcal{U}} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \right\} \leq 6\theta_0 \|\mathbf{x}_{\mathcal{I}+\mathcal{U}} - \mathbf{x}^\star\|_{\mathbf{H}^\star},$$

where the last inequality is due to $\mathbf{x}_{\mathcal{I}+\mathcal{U}} \in \mathcal{N}_\nu$ and $2\nu \leq \theta_0$. Furthermore, we can see that

$$\frac{2L}{\lambda_{\min}^{3/2}} \|\mathbf{x}_{\mathcal{I}+\mathcal{U}+1} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq \frac{2L}{\lambda_{\min}^{3/2}} \cdot 6\theta_0 \|\mathbf{x}_{\mathcal{I}+\mathcal{U}} - \mathbf{x}^\star\|_{\mathbf{H}^\star} \leq 6\theta_0^2 = (6\Psi\theta_0) \cdot \frac{\theta_0}{\Psi} \overset{(C.22)}{\leq} \theta_1.$$

Thus, (C.21) holds for $t = 0$. Suppose (C.21) holds for $t-1$ with $t \geq 1$, we prove (C.21) for t. We still characterize the difference between $\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}$ and $\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}$. We have

$$\begin{aligned}
&\|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}-\mathbf{I}\| \\
&\quad\le \|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\|+\|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\mathbf{H}^\star\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}-\mathbf{I}\| \\
&\quad\overset{\text{Lemma 12}}{\le} \|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\|+\frac{L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}.
\end{aligned} \tag{C.25}$$

For the first term on the right hand side, we have

$$\begin{aligned}
&\|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}(\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\| \\
&\quad\overset{(11)}{\le} \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}+t}\|}{\lambda_{\min}}+\frac{1}{w(\mathcal{I}+\mathcal{U}+t)} \\
&\quad\left\{\left(\sum_{j=0}^{\mathcal{I}-1}+\sum_{j=\mathcal{I}}^{\mathcal{I}+\mathcal{U}}+\sum_{j=\mathcal{I}+\mathcal{U}+1}^{\mathcal{I}+\mathcal{U}+t}\right)(w(j)-w(j-1))\|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}(\mathbf{H}_j-\mathbf{H}^\star)\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\|\right\} \\
&\quad\le \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}+t}\|}{\lambda_{\min}}+\frac{2w(\mathcal{I}-1)\kappa+w(\mathcal{I}+\mathcal{U})\nu}{w(\mathcal{I}+\mathcal{U}+t)} \\
&\qquad+\sum_{j=1}^{t}\frac{L(w(\mathcal{I}+\mathcal{U}+j)-w(\mathcal{I}+\mathcal{U}+j-1))}{\lambda_{\min}^{3/2}w(\mathcal{I}+\mathcal{U}+t)}\frac{\|\mathbf{x}_{\mathcal{I}+\mathcal{U}}-\mathbf{x}^\star\|_{\mathbf{H}^\star}}{(2\Psi)^j} \\
&\quad\le \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}+t}\|}{\lambda_{\min}}+\frac{2w(\mathcal{I}-1)\kappa+w(\mathcal{I}+\mathcal{U})\nu}{w(\mathcal{I}+\mathcal{U}+t)}+\frac{\nu}{w(\mathcal{I}+\mathcal{U}+t)}\sum_{j=1}^{t}\frac{w(\mathcal{I}+\mathcal{U}+j)}{(2\Psi)^j} \\
&\quad\le \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}+t}\|}{\lambda_{\min}}+\frac{2w(\mathcal{I}-1)\kappa+w(\mathcal{I}+\mathcal{U})\nu}{w(\mathcal{I}+\mathcal{U}+t)}+\frac{\nu w(\mathcal{I}+\mathcal{U})}{w(\mathcal{I}+\mathcal{U}+t)}\sum_{j=1}^{t}\frac{w(\mathcal{I}+\mathcal{U}+j)}{w(\mathcal{I}+\mathcal{U})(2\Psi)^j} \\
&\quad\le \frac{\|\bar{\mathbf{E}}_{\mathcal{I}+\mathcal{U}+t}\|}{\lambda_{\min}}+\frac{2w(\mathcal{I}-1)\kappa+2w(\mathcal{I}+\mathcal{U})\nu}{w(\mathcal{I}+\mathcal{U}+t)} \quad \text{(Assumption 3(v))} \\
&\quad\le \theta_t,
\end{aligned} \tag{C.26}$$

where the second inequality uses the hypothesis and the fact that the convergence rate $6\theta_t \le 1/(2\Psi)$. Thus, combining (C.25) and (C.26),

$$\|\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}\tilde{\mathbf{H}}_{\mathcal{I}+\mathcal{U}+t}\mathbf{H}_{\mathcal{I}+\mathcal{U}+t}^{-1/2}-\mathbf{I}\| \le \theta_t+\nu \overset{(C.22)}{\le} \epsilon+4\nu \overset{(26)}{\le} \frac{0.5-\beta}{1.5-\beta}.$$

Thus, the conditions of Lemma 5 are satisfied, which leads to

$$\begin{aligned}
\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star} &\le 3\left\{\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}^2+\theta_t\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}\right\} \\
&\le 6\theta_t\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}.
\end{aligned} \tag{C.27}$$

The last inequality uses the hypothesis. This shows the first result in (C.21). For the second result, we have

$$\frac{2L}{\lambda_{\min}^{3/2}}\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t+1}-\mathbf{x}^\star\|_{\mathbf{H}^\star} \overset{(C.27)}{\leq} \frac{2L}{\lambda_{\min}^{3/2}}\cdot 6\theta_t\|\mathbf{x}_{\mathcal{I}+\mathcal{U}+t}-\mathbf{x}^\star\|_{\mathbf{H}^\star}$$
$$\leq 6\theta_t^2 = (6\theta_t\Psi)\cdot\frac{\theta_t}{\Psi} \overset{(C.22)}{\leq} \theta_{t+1}.$$

This completes the induction and finishes the proof.

References

1. Agarwal, N., Bullins, B., Hazan, E.: Second-order stochastic optimization for machine learning in linear time. J. Mach. Learn. Res. **18**(1), 4148–4187 (2017)
2. Allen-Zhu, Z., Yuan, Y.: Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In: International Conference on Machine Learning, PMLR, pp. 1080–1089 (2016)
3. Bellavia, S., Krejić, N., Jerinkić, N.K.: Subsampled inexact Newton methods for minimizing large sums of convex functions. IMA J. Numer. Anal. **40**(4), 2309–2341 (2019)
4. Berahas, A.S., Bollapragada, R., Nocedal, J.: An investigation of Newton-sketch and subsampled Newton methods. Optim. Methods Softw. **35**(4), 661–680 (2020)
5. Blanchet, J., Cartis, C., Menickelly, M., Scheinberg, K.: Convergence rate analysis of a stochastic trust-region method via supermartingales. INFORMS J. Optim. **1**(2), 92–119 (2019)
6. Bollapragada, R., Byrd, R.H., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization. IMA J. Numer. Anal. **39**(2), 545–578 (2018)
7. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)
8. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
9. Bubeck, S.: Convex optimization: algorithms and complexity. Found. Trends® Mach. Learn. **8**(3–4), 231–357 (2015)
10. Byrd, R.H., Chin, G.M., Neveitt, W., Nocedal, J.: On the use of stochastic Hessian information in optimization methods for machine learning. SIAM J. Optim. **21**(3), 977–995 (2011)
11. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. Math. Program. **134**(1), 127–155 (2012)
12. Chen, R., Menickelly, M., Scheinberg, K.: Stochastic optimization using a trust-region method and random models. Math. Program. **169**(2), 447–487 (2017)
13. Clarkson, K.L., Woodruff, D.P.: Low-rank approximation and regression in input sparsity time. J. ACM **63**(6), 1–45 (2017)
14. Dennis, J.E., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**(126), 549–560 (1974)
15. Dereziński, M., Mahoney, M.W.: Distributed estimation of the inverse Hessian by determinantal averaging. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019). https://proceedings.neurips.cc/paper/2019/file/b1b20d09041289e6c3fbb81850c5da54-Paper.pdf
16. Dereziński, M., Rebrova, E.: Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. arXiv:2208.09585 (2022)
17. Dereziński, M., Bartan, B., Pilanci, M., Mahoney, M.W.: Debiasing distributed second order optimization with surrogate sketching and scaled regularization. In: Thirty-fourth Conference on Neural Information Processing Systems (2020a)
18. Dereziński, M., Liang, F.T., Liao, Z., Mahoney, M.W.: Precise expressions for random projections: low-rank approximation and randomized Newton. In: Thirty-fourth Conference on Neural Information Processing Systems (2020b)
19. Dereziński, M., Lacotte, J., Pilanci, M., Mahoney, M.W.: Newton-LESS: Sparsification without trade-offs for the sketched Newton update. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021a)

20. Dereziński, M., Liao, Z., Dobriban, E., Mahoney, M.: Sparse sketches with small inversion bias. In: Conference on Learning Theory, PMLR, pp. 1467–1510 (2021b)
21. Doikov, N., Richtárik, P., et al.: Randomized block cubic Newton method. In: International Conference on Machine Learning, PMLR, pp. 1290–1298 (2018)
22. Erdogdu, M.A., Montanari, A.: Convergence rates of sub-sampled Newton methods. In: Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2, MIT Press, Cambridge, MA, USA, NIPS'15, pp. 3052–3060 (2015)
23. Fong, D.C.L., Saunders, M.: CG versus MINRES: an empirical comparison. Sultan Qaboos Univ. J. Sci. [SQUJS] **16**(1), 44 (2012)
24. Friedlander, M.P., Schmidt, M.: Hybrid deterministic-stochastic methods for data fitting. SIAM J. Sci. Comput. **34**(3), A1380–A1405 (2012)
25. Garber, D., Hazan, E.: Fast and simple pca via convex optimization. arXiv:1509.05647 (2015)
26. Garber, D., Hazan, E., Jin, C., Musco, C., Netrapalli, P., Sidford, A., et al.: Faster eigenvector computation via shift-and-invert preconditioning. In: International Conference on Machine Learning, PMLR, pp. 2626–2634 (2016)
27. Gower, R., Kovalev, D., Lieder, F., Richtárik, P.: RSN: randomized subspace Newton. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019). https://proceedings.neurips.cc/paper/2019/file/bc6dc48b743dc5d013b1abaebd2faed2-Paper.pdf
28. Gower, R.M., Richtárik, P.: Randomized iterative methods for linear systems. SIAM J. Matrix Anal. Appl. **36**(4), 1660–1690 (2015)
29. Gross, D., Nesme, V.: Note on sampling without replacing from a finite collection of matrices. arXiv:1001.2738 (2010)
30. Islamov, R., Qian, X., Richtárik, P.: Distributed second order methods with fast rates and compressed communication. In: International Conference on Machine Learning, PMLR, pp. 4617–4628 (2021)
31. Jin, Q., Mokhtari, A.: Non-asymptotic superlinear convergence of standard quasi-Newton methods. arXiv:2003.13607 (2020)
32. Kovalev, D., Mishchenko, K., Richtárik, P.: Stochastic newton and cubic newton methods with simple local linear-quadratic rates. arXiv:1912.01597 (2019)
33. Kylasa, S., Roosta, F.F., Mahoney, M.W., Grama, A.: GPU accelerated sub-sampled Newton's method for convex classification problems. In: Proceedings of the 2019 SIAM International Conference on Data Mining, SIAM, Society for Industrial and Applied Mathematics, pp. 702–710 (2019)
34. Lacotte, J., Wang, Y., Pilanci, M.: Adaptive Newton sketch: linear-time optimization with quadratic convergence and effective Hessian dimensionality. In: Proceedings of the 38th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol. 139, pp. 5926–5936 (2021)
35. Lan, G.: First-Order and Stochastic Optimization Methods for Machine Learning. Springer, Berlin (2020)
36. Li, X., Wang, S., Zhang, Z.: Do subsampled Newton methods work for high-dimensional data? In: Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI), vol. 34, pp. 4723–4730 (2020)
37. Liu, Y., Roosta, F.: Convergence of Newton-MR under inexact Hessian information. SIAM J. Optim. **31**(1), 59–90 (2021)
38. Luo, H., Agarwal, A., Cesa-Bianchi, N., Langford, J.: Efficient second order online learning by sketching. In: Advances in Neural Information Processing Systems, pp. 902–910 (2016)
39. Ma, C., Liu, X., Wen, Z.: Globally convergent Levenberg–Marquardt method for phase retrieval. IEEE Trans. Inf. Theory **65**(4), 2343–2359 (2019)
40. Martens, J.: Deep learning via Hessian-free optimization. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, vol. 27, pp. 735–742 (2010)
41. Martens, J., Sutskever, I.: Learning recurrent neural networks with Hessian-free optimization. In: Proceedings of the 28th International Conference on International Conference on Machine Learning (2011)
42. Meng, S.Y., Vaswani, S., Laradji, I.H., Schmidt, M., Lacoste-Julien, S.: Fast and furious convergence: Stochastic second order methods under interpolation. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 1375–1386 (2020)
43. Mutný, M., Dereziński, M., Krause, A.: Convergence analysis of block coordinate algorithms with determinantal sampling. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp. 3110–3120 (2020)

44. Na, S., Mahoney, M.W.: Asymptotic convergence rate and statistical inference for stochastic sequential quadratic programming. arXiv:2205.13687 (2022)
45. Na, S., Anitescu, M., Kolar, M.: Inequality constrained stochastic nonlinear optimization via active-set sequential quadratic programming. arXiv:2109.11502 (2021)
46. Na, S., Anitescu, M., Kolar, M.: An adaptive stochastic sequential quadratic programming with differentiable exact augmented lagrangians. Math. Program. (2022). https://doi.org/10.1007/s10107-022-01846-z
47. Nocedal, J., Wright, S.J.: Numerical Optimization, Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2006)
48. Pilanci, M., Wainwright, M.J.: Newton sketch: a near linear-time optimization algorithm with linear-quadratic convergence. SIAM J. Optim. **27**(1), 205–245 (2017)
49. Qian, X., Islamov, R., Safaryan, M., Richtárik, P.: Basis matters: better communication-efficient second order methods for federated learning. In: International Conference on Artificial Intelligence and Statistics (2022)
50. Rodomanov, A., Nesterov, Y.: Greedy quasi-Newton methods with explicit superlinear convergence. SIAM J. Optim. **31**(1), 785–811 (2021)
51. Rodomanov, A., Nesterov, Y.: New results on superlinear convergence of classical quasi-Newton methods. J. Optim. Theory Appl. **188**(3), 744–769 (2021)
52. Rodomanov, A., Nesterov, Y.: Rates of superlinear convergence for classical quasi-Newton methods. Math. Program. **194**, 159–190 (2021)
53. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods. Math. Program. **174**(1–2), 293–326 (2018)
54. Safaryan, M., Islamov, R., Qian, X., Richtárik, P.: Fednl: making newton-type methods applicable to federated learning. arXiv:2106.02969 (2021)
55. Shalev-Shwartz, S.: Sdca without duality, regularization, and individual convexity. In: International Conference on Machine Learning, PMLR, pp. 747–754 (2016)
56. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning. Cambridge University Press, Cambridge (2009)
57. Tropp, J.: Freedman's inequality for matrix martingales. Electron. Commun. Probab. **16**(none), 262–270 (2011)
58. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. Comput. Math. **12**(4), 389–434 (2011)
59. Vershynin, R.: High-Dimensional Probability, vol. 47. Cambridge University Press, Cambridge (2018)
60. Wang, Z., Zhou, Y., Liang, Y., Lan, G.: Stochastic variance-reduced cubic regularization for nonconvex optimization. In: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, pp. 2731–2740 (2019)
61. Xu, P., Yang, J., Roosta-Khorasani, F., Ré, C., Mahoney, M.W.: Sub-sampled Newton methods with non-uniform sampling. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (2016)
62. Xu, P., Roosta, F., Mahoney, M.W.: Second-order optimization for non-convex machine learning: an empirical study. In: Proceedings of the 2020 SIAM International Conference on Data Mining, SIAM, Society for Industrial and Applied Mathematics, pp. 199–207 (2020)
63. Yao, Z., Xu, P., Roosta, F., Wright, S.J., Mahoney, M.W.: Inexact Newton-CG algorithms with complexity guarantees. arXiv:2109.14016 (2021)
64. Ye, H., Luo, L., Zhang, Z.: Approximate Newton methods and their local convergence. In: International Conference on Machine Learning, PMLR, pp. 3931–3939 (2017)