Adaptive Newton Sketch: Linear-time Optimization with Quadratic Convergence and Effective Hessian Dimensionality

Jonathan Lacotte, Yifei Wang, Mert Pilanci
Stanford University, Electrical Engineering

Abstract

We propose a randomized algorithm with quadratic convergence rate for convex optimization problems with a self-concordant, composite, strongly convex objective function. Our method is based on performing an approximate Newton step using a random projection of the Hessian. Our first contribution is to show that, at each iteration, the embedding dimension (or sketch size) can be as small as the effective dimension of the Hessian matrix. Leveraging this novel fundamental result, we design an algorithm with a sketch size proportional to the effective dimension and which exhibits a quadratic rate of convergence. This result dramatically improves on the classical linear-quadratic convergence rates of state-of-the-art sub-sampled Newton methods. However, in most practical cases, the effective dimension is not known beforehand, and this raises the question of how to pick a sketch size as small as the effective dimension while preserving a quadratic convergence rate. Our second and main contribution is thus to propose an adaptive sketch size algorithm with quadratic convergence rate and which does not require prior knowledge or estimation of the effective dimension: at each iteration, it starts with a small sketch size, and increases it until quadratic progress is achieved. Importantly, we show that the embedding dimension remains proportional to the effective dimension throughout the entire path and that our method achieves state-of-the-art computational complexity for solving convex optimization programs with a strongly convex component. We discuss and illustrate applications to linear and quadratic programming, as well as logistic regression and other generalized linear models.

1 Introduction

We consider a composite optimization problem of the form

$$x^* := \operatorname*{argmin}_{x \in \mathbb{R}^d} \{f(x) := f_0(x) + g(x)\} , \tag{1}$$

where $f_0, g : \mathbb{R}^d \to \overline{\mathbb{R}}$ are both closed, twice differentiable convex functions. Here, we denote $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ and by $\mathbf{dom}\, f$ the domain of f. We are interested in the structured setting where forming the Hessian matrix $\nabla^2 f_0(x)$ is prohibitively expensive, but we have available at small computational cost a Hessian matrix square-root $\nabla^2 f_0(x)^{1/2}$, that is, a matrix $\nabla^2 f_0(x)^{1/2}$ of dimensions $n \times d$ such that $(\nabla^2 f_0(x)^{1/2})^\top \nabla^2 f_0(x)^{1/2} = \nabla^2 f_0(x)$ for some integer $n \geqslant d$, and n eventually very large. Moreover, we assume the function g to be μ-strongly convex, i.e., $\nabla^2 g(x) \succeq \mu I_d$.

Large-scale optimization problems of this form are becoming ever more common in applications, due to the increasing dimensionality of data (e.g., genomics, medicine, high-dimensional models). Typically, the function f_0 may represent an objective value we aim to minimize over a convex set $\mathcal{C} \subseteq \mathbb{R}^d$, that is, we aim to solve $\min_{x \in \mathcal{C}} f_0(x)$. A common practice to turn this constrained optimization problem into an unconstrained one is to add to the objective function a *penalty* or barrier function $g(x)$ which encodes $\mathcal{C}$ (e.g., logarithmic barrier functions for polyhedral constraints or ℓ_p-norm regularization for ℓ_p-ball constraints). In many cases of practical interest, a matrix square-root $\nabla^2 f_0(x)^{1/2}$ can be computed efficiently. For instance, in the broad context of empirical risk minimization, the function f_0 has the separable form $f_0(x) = \sum_{i=1}^m \ell_i(a_i^\top x)$ where the functions ℓ_i are twice-differentiable and convex. In this case, a suitable Hessian matrix square root is

given by the $n \times d$ matrix $\nabla^2 f_0(x)^{1/2} := \mathbf{diag}(\ell_i''(a_i^\top x)^{1/2}) \cdot A$. On the other hand, we assume that the Hessian of the function g is well-structured, so that its computation is relatively cheap in comparison to that of f_0. For instance, if the constraint set is the unit simplex (i.e., $x \geqslant 0$ and $\mathbf{1}^\top x \leqslant 1$), then the Hessian of the associated logarithmic barrier function is a diagonal matrix plus a rank one matrix. Other examples include problems for which g has a separable structure such as typical regularizers for ill-posed inverse problems (e.g., graph regularization $g(x) = \frac{1}{2}\sum_{i,j\in E}(x_i - x_j)^2$, ℓ_p-norms with $p > 1$ or approximations of ℓ_1-norm).

Second-order methods such as the Newton's method enjoy superior convergence in both theory and practice compared to first-order methods, that is, quadratic convergence rate versus $1/T^2$ for accelerated gradient descent. A common issue in first-order methods is the tuning of step size [3], whose optimal choice depends on the strong convexity and smoothness of the underlying problem. In contrast, whenever the objective function f is *self-concordant*, then Newton's method has the appealing property of being invariant to rescaling and coordinate transformations, is independent of problem-dependent parameters, and thus needs little or no tuning of algorithmic hyperparameters. More precisely, we recall that, given a current iterate x, the standard Newton's method computes the Hessian matrix $H(x)$ and the descent direction v_{ne} defined as

$$H(x) := \nabla^2 f_0(x) + \nabla^2 g(x)\,, \tag{2}$$

$$v_{\mathrm{ne}} := -H(x)^{-1}\nabla f(x)\,. \tag{3}$$

Given a step size $s > 0$, it then uses the update

$$x_{\mathrm{ne}} := x + s \cdot v_{\mathrm{ne}}\,. \tag{4}$$

Despite these advantages, Newton's method requires, at each iteration, forming and solving the high-dimensional linear system $H(x)v_{\mathrm{ne}} = -\nabla f(x)$, which has complexity scaling as $\mathcal{O}(nd^2)$, and this becomes prohibitive in large-scale settings. To address this numerical challenge, a multitude of different approximations to Newton's method have been proposed in the literature. Quasi-Newton methods (e.g, DFP, BFGS and their limited memory versions [30]) are computationally cheaper, but their convergence guarantees require stronger assumptions and are typically much weaker than those of Newtons's method. On the other hand, random projections are an effective way of performing dimensionality reduction [36, 23, 17], and many random projection (or *sketching*) based algorithms were designed to reduce the cost of solving the linear Newton system. For instance, the respective methods in [18] and [21] embed the optimization variable into a lower dimensional subspace, so that solving the Newton system becomes cheaper; [32] propose to solve an approximate Newton system based on random principal sub-matrices of a global upper bound on the Hessian; [16] address a common setting, that of block-separable convex optimization problems, and propose a method combining the ideas of randomized coordinate descent with cubic regularization [28, 29].

Our work builds specifically on a generic method, that is, the Newton sketch [31], which is based on a structured random embedding of the Hessian matrix $H(x)$. Formally, given a sketch size m such that $m \ll n$ and an embedding matrix $S \in \mathbb{R}^{m\times n}$ to be precised, the Newton sketch computes the approximate Hessian $H_S(x)$ and the approximate descent direction v_{nsk} defined as

$$H_S(x) := (\nabla^2 f_0(x)^{\frac{1}{2}})^\top S^\top S \nabla^2 f_0(x)^{\frac{1}{2}} + \nabla^2 g(x)\,, \tag{5}$$

$$v_{\mathrm{nsk}} := -H_S(x)^{-1}\nabla f(x)\,. \tag{6}$$

Given a step size $s > 0$, it then uses the update

$$x_{\mathrm{nsk}} := x + s \cdot v_{\mathrm{nsk}}\,. \tag{7}$$

For classical embeddings (e.g., sub-Gaussian, randomized orthogonal systems), it has been shown by [31] that, in general, a sketch size $m \asymp d$ is sufficient for the Newton sketch to achieve a linear-quadratic convergence rate with high probability (w.h.p.).

Contributions. Our first key contribution is to show that, under the assumption that g is μ-strongly convex, the scaling $m \asymp \overline{d}_\mu \log(\overline{d}_\mu)/\delta$ is sufficient for the Newton sketch to achieve a δ-accurate solution at a *quadratic*

convergence rate with high probability. More generally, we show that convergence is geometric provided that m scales appropriately in terms of $\overline{d}_\mu$. Here, the critical quantity $\overline{d}_\mu$ is the effective (Hessian) dimension, defined as

$$\overline{d}_\mu := \sup_{x \in \mathcal{S}(x_0)} d_\mu(x)\,, \tag{8}$$

where x_0 is the initial point of our algorithm, $\mathcal{S}(x_0)$ is the sublevel set of f at x_0, and

$$d_\mu(x) := \operatorname{trace}(\nabla^2 f_0(x)(\nabla^2 f_0(x) + \mu I_d)^{-1}) \tag{9}$$

is the *local* effective dimension. Importantly, it always holds that $d_\mu(x) \leqslant \overline{d}_\mu \leqslant \min\{n, d\} = d$. In many applications, the effective dimension is substantially smaller than the ambient dimension d [6, 2, 41]. However, in order to pick m in terms of $\overline{d}_\mu$ which is usually unknown and then achieve computational and memory space savings, it is necessary to estimate $\overline{d}_\mu$. There exist randomized techniques for precise estimation of $d_\mu(x)$, but they provably work under stringent assumptions, e.g., $d_\mu(x)$ very small (e.g., see Theorem 60 in [4]). In the context of ridge regression, [20] proposed a sketching-based method with adaptive (time-varying) sketch size scaling as the effective dimension, and without prior knowledge or estimation of it. Starting with a small sketch size, it checks at each iteration whether enough progress is achieved by the update. If not, it doubles the sketch size. The time and memory complexities of this method to return a *certified* δ-accurate solution w.h.p. scale in terms of the effective dimension, i.e., it takes time $\mathcal{O}(nd\log^2(\overline{d}_\mu)\log(d/\delta))$ with a sketch size $m \lesssim \overline{d}_\mu \log(\overline{d}_\mu)$ for large values of n. This significantly improves on usual standard randomized pre-conditioning methods [33, 5, 24] which require $m \gtrsim d$.

In a vein similar to this adaptive ridge regression solver, our second key contribution is to propose an adaptive sketch size version of the effective dimension Newton sketch. Importantly, we prove that the adaptive sketch size scales in terms of $\overline{d}_\mu$. Furthermore, our adaptive method offers the possibility to the user to choose the convergence rate, from linear to quadratic.

Other related works. Recent studies in the literature on randomized second-order and Sub-sampled Newton methods [11, 9, 34, 8] show that picking an embedding dimension proportional to d and possibly smaller than d under certain conditions do work empirically in many settings [40, 39, 37]. The recent work by [22] provides a more precise understanding of these phenomena. In the context of empirical risk minimization with ℓ_2^2-regularization, they show that the subsampled Newton method with $m \asymp \overline{d}_\mu$ data points is enough to guarantee convergence. However, differently from our work, their method needs to estimate the effective dimension at each iteration. Furthermore, their convergence guarantees severely depend on the condition number of the problem (e.g., see their Theorems 1 and 2), whereas our results are independent of condition numbers and only involve the relevant dimensions of the problem $(n, d, \overline{d}_\mu)$ and the target accuracy. Besides effective dimension based sampling, sketching-based methods are used in the context of distributed optimization where due to stringent memory and/or communication constraints, reducing the number of iterations and/or the size of second-order information is critical [35, 14, 7].

1.1 Notations and background

A closed convex function $\varphi : \mathbb{R}^d \to \overline{\mathbb{R}}$ is *self-concordant* if $|\varphi'''(x)| \leqslant 2 \cdot (\varphi''(x))^{3/2}$. This definition extends to a closed convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ by imposing this requirement on the univariate functions $\varphi_{x,y}(t) := f(x + ty)$ for all choices of x, y in the domain of f. Self-concordance is a typical assumption for the analysis of the classical Newton's method, in order to obtain convergence results which are independent of unknown problem parameters (e.g., strong convexity, smoothness or Lipschitz constants; see the books by [27] or [10] for further background), and this encompasses many widely used functions in practice, e.g., linear, quadratic, negative logarithm. Hence, in this work, we assume that *f_0 and g are self-concordant functions.*

The choice of the sketching matrix $S \in \mathbb{R}^{m \times n}$ is critical for statistical and computational performances. The well-structured subsampled randomized Hadamard transform (SRHT) [1] usually serves as a reference for

comparing sketching algorithms thanks to its strong subspace embedding properties [23, 17, 15, 19] and fast sketching time $\mathcal{O}(nd \log m)$ compared to the classical sketching cost $\mathcal{O}(ndm)$ of sub-Gaussian embeddings. Another typical choice is the sparse Johnson-Lindenstrauss transform (SJLT) [25, 38] with, for instance, one non-zero entry per column. With $A \in \mathbb{R}^{n \times d}$, a sketch SA is then much faster to compute (it takes time $\mathcal{O}(\mathbf{nnz}(A))$) at the expense of weaker subspace embedding properties.

1.2 Organization of the paper

In Section 2, we introduce critical quantities and preliminary results for both the implementation of our algorithms and their analysis. We show that the approximate Newton direction $v_{\rm nsk}$ is close to the exact one $v_{\rm ne}$, provided that the sketch size scales in terms of $\overline{d}_\mu$. In Section 3, we formally introduce our (non-adaptive) effective dimension Newton sketch algorithm (see Algorithm 1), and we present several relevant applications. Assuming knowledge of $\overline{d}_\mu$, we prove that its convergence rate is geometric. In Section 4, we introduce an adaptive version of Algorithm 1 (see Algorithm 2): importantly, it does not require knowledge of $\overline{d}_\mu$, but still guarantees geometric convergence as well as low memory complexity in terms of $\overline{d}_\mu$. We summarize our complexity guarantees in Table 1 and compare to standard first- and second-order methods and to the original Newton sketch algorithm [31] whose implementation and guarantees are agnostic to the effective dimension of the problem. Finally, we show in Section 5 the empirical benefits of our adaptive method, compared to several standard optimization baselines.

2 Preliminaries

Critical to our algorithms and their analysis are the Newton and approximate Newton decrements, defined as

$$\lambda_f(x) := \left(\nabla f(x)^\top H(x)^{-1} \nabla f(x)\right)^{\frac{1}{2}}, \tag{10}$$

$$\widetilde{\lambda}_f(x) := \left(\nabla f(x)^\top H_S(x)^{-1} \nabla f(x)\right)^{\frac{1}{2}}. \tag{11}$$

Importantly, for a self-concordant function f, the optimality gap at any point $x \in \mathbf{dom}\, f$ is bounded in terms of the Newton decrement as

$$f(x) - f(x^*) \leqslant \lambda_f(x)^2 . \tag{12}$$

Due to the expensive cost of computing the Newton decrement $\lambda_f(x)$ as opposed to $\widetilde{\lambda}_f(x)$, we will aim to characterize, w.h.p. over the randomness of the sketching matrix, similar optimality bounds and properties with $\widetilde{\lambda}_f(x)$.

Given $x \in \mathbf{dom}\, f$, a sketch size $m \geqslant 1$, a random embedding $S \in \mathbb{R}^{m \times d}$ and a sampling precision parameter $\varepsilon > 0$, we consider the following probability event which is critical to our convergence guarantees,

$$\boxed{\mathcal{E}_{x,m,\varepsilon} := \left\{(1 - \frac{\varepsilon}{2}) I_d \preceq C_S \preceq (1 + \frac{\varepsilon}{2}) I_d\right\}}, \tag{13}$$

where $C_S := H^{-\frac{1}{2}} H_S H^{-\frac{1}{2}}$, $H \equiv H(x)$ and $H_S \equiv H_S(x)$. In words, when $\mathcal{E}_{x,m,\varepsilon}$ holds true, the matrix $H^{-1/2} H_S H^{-1/2}$ is a close approximation of the identity, i.e., $H^{-1/2} H_S H^{-1/2} \approx H^{-1/2} H H^{-1/2}$. The next result bounds the probability for this event to hold for different choices of the sketching matrix.

Lemma 1. *Let $\varepsilon \in (0, 1/4)$ and $p \in (0, 1/2)$. It holds that $\mathbb{P}(\mathcal{E}_{x,\varepsilon,m}) \geqslant 1 - p$, provided that*

$$m = \Omega(d_\mu(x)^2/(\varepsilon^2 p)) \qquad \textit{for the SJLT with single nonzero element in each column}, \tag{14}$$

$$m = \Omega((d_\mu(x) + \log(1/\varepsilon p) \log(d_\mu(x)/p))/\varepsilon^2) \qquad \textit{for the SRHT}. \tag{15}$$

We show next that conditional on $\mathcal{E}_{x,m,\varepsilon}$, the approximate Newton decrement $\widetilde{\lambda}_f(x)$ is close to $\lambda_f(x)$, as well as the approximate Newton direction $v_{\rm nsk}$ to the exact one $v_{\rm ne}$.

Theorem 1 (Closeness of Newton decrements). *Let $\varepsilon \in (0, 1/4)$. Conditional on the event $\mathcal{E}_{x,m,\varepsilon}$, it holds that*

$$\|v_{\mathrm{ne}} - v_{\mathrm{nsk}}\|_{H(x)} \leqslant \varepsilon \cdot \|v_{\mathrm{ne}}\|_{H(x)}\,, \tag{16}$$

$$\sqrt{1-\varepsilon} \cdot \lambda_f(x) \leqslant \widetilde{\lambda}_f(x) \leqslant \sqrt{1+\varepsilon} \cdot \lambda_f(x)\,. \tag{17}$$

Given $\varepsilon \in (0, 1/4)$, we introduce positive parameters a, b such that $1 - \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2 \geqslant a$, which we use for backtracking line-search (see Algorithm 1 for details). Furthermore, we define the parameters

$$\eta := \frac{1}{8} \cdot \big(1 - \frac{1}{2}\big(\frac{1+\varepsilon}{1-\varepsilon}\big)^2 - a\big) / \big(\frac{1+\varepsilon}{1-\varepsilon}\big)^3\,,$$

$$\nu := ab \cdot \frac{\eta^2}{1 + \frac{1+\varepsilon}{1-\varepsilon} \cdot \eta}\,.$$

The next results aim to describe the empirical behavior of our methods. As for the classical Newton's method, we distinguish two phases. The algorithm follows a first phase with constant additive decrease in objective value. In a second phase, it converges faster, i.e., the Newton decrement converges to zero at a geometric rate up to quadratic for an appropriate choice of the hyperparameters.

Lemma 2 (First phase decrement). *Let $\varepsilon \in (0, 1/4)$. Suppose that $\mathcal{E}_{x,m,\varepsilon}$ holds true and that $\widetilde{\lambda}_f(x) > \eta$. Then, we have that*

$$f(x_{\mathrm{nsk}}) - f(x) \leqslant -\nu\,. \tag{18}$$

We introduce the following numerical function which will prove to be useful to characterize the rate of convergence of our algorithms,

$$\alpha(\tau) := 0.57 + \frac{16^\tau}{15}\,. \tag{19}$$

It is easy to verify that $\alpha(\tau)^{1/\tau} \leqslant 2$ for $\tau \in (0, 1]$ and $\alpha(0) \leqslant \frac{16}{25}$.

Lemma 3 (Second phase decrement). *Let $x \in \mathbf{dom}\, f$, $\tau \in [0, 1]$ and $\varepsilon \in (0, 1/4)$. Set $\varepsilon' = \varepsilon \cdot \min\{1, \lambda_f(x)^\tau\}$. We assume that the event $\mathcal{E}_{x,m,\varepsilon'}$ holds and that $\widetilde{\lambda}_f(x) \leqslant \eta$. Then, we have*

$$\lambda_f(x_{\mathrm{nsk}}) \leqslant \alpha(\tau) \cdot \lambda_f(x)^{1+\tau}\,. \tag{20}$$

Consequently, the progress is geometric for any $\tau \in (0, 1]$, i.e.,

$$\alpha(\tau)^{1/\tau} \cdot \lambda_f(x_{\mathrm{nsk}}) \leqslant \Big(\alpha(\tau)^{1/\tau} \cdot \lambda_f(x)\Big)^{1+\tau}\,. \tag{21}$$

On the other hand, the progress is linear for $\tau = 0$, i.e.,

$$\lambda_f(x_{\mathrm{nsk}}) \leqslant \frac{16}{25} \cdot \lambda_f(x)\,. \tag{22}$$

We conclude this section with a simple technical lemma which characterizes a sufficient number of iterations before termination, under geometric convergence.

Lemma 4 (Geometric convergence and sufficient iteration number). *Let $\delta \in (0, 1)$, $\alpha > 0$, $\tau \in (0, 1]$, and $\{\beta_t\}_{t \geqslant 0}$ be a sequence of positive numbers such that $\beta_0 \leqslant \eta$, $\eta\alpha^{1/\tau} < 1$, $\sqrt{\delta}\alpha^{1/\tau} < 1$ and $\alpha^{1/\tau}\beta_{t+1} \leqslant (\alpha^{1/\tau}\beta_t)^{1+\tau}$ for all $t \geqslant 0$. Then, it holds that $\beta_t \leqslant \sqrt{\delta}$ for any $t \geqslant T_{\tau,\alpha,\delta}$ where*

$$T_{\tau,\alpha,\delta} := \lceil \frac{1}{\log(1+\tau)} \cdot \log\left(\frac{1 + \frac{\tau \log(1/\delta)}{2\log(1/\alpha)}}{1 + \frac{\tau \log(1/\eta)}{\log(1/\alpha)}}\right) \rceil\,. \tag{23}$$

Throughout this work, we will use the shorthand

$$\boxed{T_{\tau,\delta} \equiv T_{\tau,\alpha(\tau),\delta}}\,. \tag{24}$$

Note in particular that $T_{\tau,\delta} = \mathcal{O}(\log(\tau \log(1/\delta)))$ for small δ. Further, it holds that $\lim_{\tau\to 0} T_{\tau,\delta} \leqslant \lceil \frac{\log(1/\delta)}{\log(25/16)} \rceil$, which corresponds to the classical complexity of linear convergence with rate $16/25$.

3 Effective dimension Newton sketch

We formally introduce our effective dimension Newton sketch method in Algorithm 1. Algorithm 1 takes as

Algorithm 1: Effective dimension Newton sketch

Require: Initial point $x_0 \in \mathbf{dom} f$, threshold sketch sizes $\overline{m}_1$ and $\overline{m}_2$, initial sketch size $m_0 = \overline{m}_1$, line-search parameters (a, b), target accuracy $\delta > 0$, convergence rate parameter $\tau \in [0, 1]$ and sampling precision parameter $\varepsilon = 1/8$.

1: **for** $t = 0, \ldots$ **do**
2: Sample an $m_t \times n$ embedding S_t independent of $\{S_j\}_{j=0}^{t-1}$. Compute v_{nsk} and $\widetilde{\lambda}_f(x_t)$ based on S_t.
3: **if** $\widetilde{\lambda}_f(x_t)^2 \leqslant \frac{3}{4}\delta$ **then return** x_t.
4: Starting at $s = 1$: **while** $f(x_t + s v_{\mathrm{nsk}}) > f(x_t) + as\nabla f(x_t)^\top v_{\mathrm{nsk}}, \quad s \leftarrow bs$.
5: Update $x_{t+1} \leftarrow x_t + s \cdot v_{\mathrm{nsk}}$.
6: If $\widetilde{\lambda}_f(x_t) > \eta$, set $m_{t+1} = \overline{m}_1$. Otherwise, set $m_{t+1} = \overline{m}_2$.
7: **end for**

inputs the phase 1 and phase 2 sketch sizes $\overline{m}_1$ and $\overline{m}_2$. As we will see in Theorem 2, sufficient values for $\overline{m}_1$ and $\overline{m}_2$ to guarantee convergence both depend on the effective dimension $\overline{d}_\mu$. Here and only for Algorithm 1, we make the idealized assumption that the quantity $\overline{d}_\mu$ is known. In contrast, we introduce in Section 4 an adaptive method that does not require knowledge of $\overline{d}_\mu$.

Theorem 2 (Geometric convergence guarantees of the Newton sketch)**.** *Let $\tau \in [0, 1]$, $\delta \in (0, 1/2)$ and $p_0 \in (0, 1/2)$. Set $\varepsilon = 1/8$. Then, the total number of iterations T_f and the total time complexity $\mathcal{C}$ for obtaining a δ-approximate solution $\widetilde{x}$ in function value (i.e., $f(\widetilde{x}) - f(x^*) \leqslant \delta$) via Algorithm 1 satisfy*

$$T_f \leqslant \overline{T} := \frac{f(x_0) - f(x^*)}{\nu} + T_{\tau,\frac{3}{8}\delta} + 1\,, \tag{25}$$

$$\mathcal{C} = \mathcal{O}\big(\overline{m}_2^2 d + nd \log \overline{m}_2\big)\,\overline{T}\,, \tag{26}$$

with probability at least $1 - p_0$, provided that

$$\overline{m}_1 \gtrsim \overline{d}_\mu + \log(\frac{\overline{T}}{p_0}) \log(\frac{\overline{d}_\mu \overline{T}}{p_0})\,, \qquad \textit{and} \qquad \overline{m}_2 \gtrsim \delta^{-\tau} \left(\overline{d}_\mu + \log(\frac{\overline{T}}{p_0 \delta^{\tau/2}}) \log(\frac{\overline{d}_\mu \overline{T}}{p_0}) \right)\,. \tag{27}$$

for the SRHT, whereas for the SJLT, it is sufficient to have

$$\overline{m}_1 \gtrsim \frac{\overline{d}_\mu^2 \overline{T}}{p_0}\,, \qquad \textit{and} \qquad \overline{m}_2 \gtrsim \frac{\overline{d}_\mu^2 \overline{T}}{\delta^\tau p_0}\,. \tag{28}$$

We draw some immediate consequences of Theorem 2, which will be useful for further discussions and comparisons of our complexity guarantees in Section 4.1. With the SRHT, consider the quadratic convergence case, i.e., $\tau = 1$. We pick a failure probability $p_0 \asymp \frac{1}{\overline{d}_\mu}$, and sketch sizes $\overline{m}_1 \asymp \overline{d}_\mu$ and $\overline{m}_2 \asymp \frac{\overline{d}_\mu \log(\overline{d}_\mu/\delta)}{\delta}$. We observe quadratic convergence with $T_f = \mathcal{O}(\log\log(\frac{1}{\delta}))$ iterations. Further, assuming that the sample size n

is large enough for the sketching cost $\mathcal{O}(nd \log \overline{m})$ to dominate the cost $\mathcal{O}(\overline{m}^2 d)$ of solving the randomized Newton system, i.e., $n \gtrsim \frac{\overline{d}_\mu^2 \log(\overline{d}_\mu/\delta)}{\delta^2}$, then the total complexity results in

$$\mathcal{C} = \mathcal{O}\big(nd \log\big(\frac{\overline{d}_\mu}{\delta}\big) \log\log(\frac{1}{\delta})\big). \tag{29}$$

Similarly, we consider the linear convergence case, i.e., $\tau = 0$. For simplicity, suppose that $\overline{d}_\mu \gtrsim \log\log(1/\delta)$. We pick $p_0 \asymp \frac{1}{\overline{d}_\mu}$, and sketch sizes $\overline{m}_1 \asymp \overline{m}_2 \asymp \overline{d}_\mu$. We observe linear convergence with $T_f = \mathcal{O}(\log \frac{1}{\delta})$ iterations. Assuming again that the sample size n is large enough for the sketching cost to dominate the cost of solving the randomized Newton system, i.e., $n \gtrsim \overline{d}_\mu^2 / \log(\overline{d}_\mu)$, we obtain the total time complexity

$$\mathcal{C} = \mathcal{O}\bigg(nd \log(\overline{d}_\mu) \log(\frac{1}{\delta})\bigg) \,. \tag{30}$$

We proceed with a similar discussion for the SJLT at the end of the proof of Theorem 2 deferred to the Appendix.

3.1 Some applications of the effective dimension Newton sketch

We discuss various concrete instantiations of the optimization problem (1) where the function g satisfies μ-strong convexity and for which forming the partially sketched Hessian $H_S(x)$ is amenable to fast computation.

Example 1 (Ridge regression)**.** *We consider the optimization problem*

$$\min_{x \in \mathbb{R}^d} \bigg\{ f(x) := \frac{1}{2}\|Ax - b\|_2^2 + \frac{\mu}{2}\|x\|_2^2 \bigg\} \tag{31}$$

where $A \in \mathbb{R}^{n \times d}$ with $n \geqslant d$ and whose solution is given in closed-form by $x^* = (A^\top A + \mu I_d)^{-1} A^\top b$. Direct methods yield the exact solution in time $\mathcal{O}(nd^2)$, whereas first-order methods (e.g., conjugate gradient method) yield an δ-approximate solution in time $\mathcal{O}(\sqrt{\kappa} nd \log(1/\delta))$, where κ is the condition number of A. Randomized pre-conditioning and sketching methods can improve on this complexity (see Section 4.1 for further details). Here, our setting for the Newton sketch applies with $f_0(x) = \frac{1}{2}\|Ax - b\|_2^2$ (whose square-root Hessian is A) and $g(x) = \frac{\mu}{2}\|x\|_2^2$ which is μ-strongly convex.

Example 2 (Portfolio optimization)**.** *The optimization problem takes the form*

$$\min_{x \geqslant 0,\, \sum_{i=1}^d x_j \leqslant 1} \big\{ f_0(x) := -r^\top x + \alpha \, \langle x, \Sigma x \rangle \big\} \,, \tag{32}$$

where $\Sigma = A^\top A$ is an empirical covariance matrix based on the data $A \in \mathbb{R}^{n \times d}$, with $n \geqslant d$. Using the barrier method, we need to solve its penalized version $\min_{x \in \mathbb{R}^d}\{f_0(x) + g(x)\}$, where $g(x) := -\mu \cdot \sum_{i=1}^d \log(x_i) - \mu \cdot \log(1 - \langle \mathbf{1}, x \rangle)$. We clearly have the Hessian square-root $\nabla^2 f_0(x)^{1/2} = \sqrt{\alpha} A$. Further, g is μ-strongly convex over its domain: indeed, note that for $0 < x_i < 1$, the Hessian of g is $\mu \, \mathbf{diag}(x_i^2)^{-1} + \mu \mathbf{1}\mathbf{1}^\top$ and the first term satisfies $\mu \, \mathbf{diag}(x_i^2)^{-1} \succeq \mu I_d$.

Example 3 (Solving Lasso via its dual)**.** *Given $A \in \mathbb{R}^{n \times d}$ with $d \gg n$, the dual Lasso problem takes the form*

$$\max_{\|A^T x\|_\infty \leqslant \lambda} \bigg\{ -\frac{1}{2}\|y - x\|_2^2 \bigg\} \,. \tag{33}$$

Applying the logarithmic barrier method, one needs to solve a sequence of problems of the form $\min_{x \in \mathbb{R}^n}\{f_0(x) + g(x)\}$ where $g(x) := \frac{\mu}{2}\|y - x\|_2^2$, $f_0(x) := -\sum_{j=1}^d \log(\lambda - \langle a_j, x \rangle) - \sum_{j=1}^d \log(\lambda + \langle a_j, x \rangle)$ and a_j is the j-th column of A. This form is amenable to the Newton sketch: a square-root of $\nabla^2 f_0(x)$ is given by $\nabla^2 f_0(x)^{1/2} = \mathbf{diag}(|\lambda - \langle a_j, x \rangle|^{-1} + |\lambda + \langle a_j, x \rangle|^{-1}) \cdot A^T$, and the function $g(x)$ is μ-strongly convex.

Example 4 (Regularized logistic regression with $n \gg d$)**.** *We consider data points $\{(a_i, y_i)\}_{i=1}^n$ where each a_i is a d-dimensional feature vector with binary response $y_i \in \{\pm 1\}$. We aim to find a linear classifier through regularized logistic regression, that is,*

$$\min_{x\in\mathbb{R}^d} \left\{ \sum_{i=1}^n \log\left(1 + e^{-y_i a_i^\top x}\right) + \frac{\mu}{2}\|x\|_2^2 \right\}. \tag{34}$$

Setting $f_0(x) = \sum_{i=1}^n \log\left(1 + e^{-y_i a_i^\top x}\right)$ and $g(x) = \frac{\mu}{2}\|x\|_2^2$, we have that $\nabla^2 f_0(x)^{\frac{1}{2}} = \mathbf{diag}(h)A$ where the i-th coefficient of $h \in \mathbb{R}^n$ is given by $h_i = \frac{e^{y_i a_i^\top x/2}}{1+e^{y_i a_i^\top x}}$. More generally, empirical risk minimization with generalized linear models yields a Hessian square-root of the form 'diagonal times data matrix A'.

Example 5 (Projection onto polyhedra)**.** *Given $v \in \mathbb{R}^d$, $A \in \mathbb{R}^{n\times d}$ with $n \gg d$ and $b \in \mathbb{R}^n$ such that there exists $x_0 \in \mathbb{R}^d$ that satisfies $Ax_0 < b$, we aim to solve the optimization problem*

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|x - v\|_2^2, \tag{35}$$

$$s.t.\ \ Ax \leqslant b. \tag{36}$$

Applying a barrier method, one needs to solve a sequence of optimization problems of the form $\min_x f_0(x)+g(x)$, where $f_0(x) := -\sum_{i=1}^n \log(b_i - a_i^\top x)$ and $g(x) = \frac{\mu}{2}\|x-v\|_2^2$. Clearly, g is μ-strongly convex, and a square-root of $\nabla^2 f_0(x)$ is given by $\mathbf{diag}(|b_i - a_i^\top x|^{-1})A$.

4 Adaptive Newton Sketch with effective dimensionality

We turn to the adaptive version of Algorithm 1, which starts with small sketch size and does not require knowledge or estimation of the effective dimension $\overline{d}_\mu$. Importantly, our method is guaranteed to converge at a tunable geometric rate, and with a sketch size scaling in terms of $\overline{d}_\mu$.

For $\tau \in [0, 1]$ and $\varepsilon \in (0, 1/4)$, we set

$$\alpha(\tau, \varepsilon) := \frac{(1+\varepsilon)^{\frac{1}{2}}}{(1-\varepsilon)^{\frac{1+\tau}{2}}} \cdot \alpha(\tau), \tag{37}$$

and we will consider in this section the sufficient number of iterations $T_{\tau,\alpha(\tau,\varepsilon),\frac{\delta}{d}}$ as defined in Lemma 4 for $\alpha = \alpha(\tau, \varepsilon)$. Our adaptive method is formally described in Algorithm 2. It starts each iteration by checking whether $\widetilde{\lambda}_f(x_t) > \eta$. If so, assuming the sketch size m_t large enough, we have w.h.p. by Lemma 2 that $f(x_{\rm nsk}) - f(x) \leqslant -\nu$ and we set $x_{t+1} = x_{\rm nsk}$. Otherwise, if $\widetilde{\lambda}_f(x_t) \leqslant \eta$, we have w.h.p. by Lemma 3 a condition similar to $\widetilde{\lambda}_f(x_{\rm nsk}) \leqslant \alpha(\tau, \varepsilon)(\widetilde{\lambda}_f(x_t))^{1+\tau}$, in which case we set $x_{t+1} = x_{\rm nsk}$. If none of the above events happen, we increase the sketch size by a factor 2. On the other hand, if the sketch size is not large enough for the guarantees of Lemmas 2 and 3 to hold w.h.p., then either the algorithm terminates with a potentially small sketch size, or, the sketch size must at some point become large enough due to the doubling trick.

Note that if Algorithm 2 terminates, then it returns an iterate x such that $\widetilde{\lambda}_f(x)^2 \leqslant \frac{\delta}{d}$. We prove next (see Lemma 5) that this termination condition implies the δ-approximation guarantee, i.e., $f(x) - f(x^*) \leqslant \delta$ w.h.p., provided that the *initial* sketch size is large enough, and regardless of the final sketch size.

Lemma 5 (Termination condition)**.** *Let $\delta \in (0, 1/2)$ and $p \in (0, 1/2)$, and suppose that Algorithm 2 returns x. Then, it holds that $f(x) - f(x^*) \leqslant \delta$ with probability at least $1 - p$ provided that*

$$m_0 \gtrsim \log^2(1/p) \quad \textit{for the SRHT}, \qquad \textit{and} \qquad m_0 \gtrsim 1/p \quad \textit{for the SJLT}. \tag{38}$$

Algorithm 2: Adaptive effective dimension Newton sketch

Require: Initial point $x_0 \in \mathbf{dom} f$, initial sketch size $m_0 = \overline{m}_0$, line-search parameters (a, b), target accuracy $\delta \in (0, 1/2)$, convergence rate parameter $\tau \in [0, 1]$ and sampling precision parameter $\varepsilon = 1/8$.

for $t = 0, \dots$ **do**
  Sample $S_t \in \mathbb{R}^{m_t \times n}$ independent of $S_{t-1}, \dots, S_0$.
  Compute v_{nsk} and $\widetilde{\lambda}_f(x_t)$ based on S_t.
  if $\widetilde{\lambda}_f(x_t)^2 \leqslant \frac{\delta}{d}$ then **return** x_t.
  Find step size s with backtracking line search, and set $x_{\mathrm{nsk}} = x_t + s v_{\mathrm{nsk}}$.
  if $\widetilde{\lambda}_f(x_t) > \eta$ **then**
    if $f(x_{\mathrm{nsk}}) - f(x) \leqslant -\nu$ **then**
      Set $x_{t+1} = x_{\mathrm{nsk}}$ and $m_{t+1} = m_t$.
    else
      Set $x_{t+1} = x_t$ and $m_{t+1} = 2m_t$.
    end if
  else
    Sample $S^+ \in \mathbb{R}^{m_t \times n}$ independent of $S_t, \dots, S_0$.
    Compute $v^+ = -H_{S^+}^{-1} \nabla f(x_{\mathrm{nsk}})$ and $\widetilde{\lambda}_f(x_{\mathrm{nsk}}) = (-\langle \nabla f(x_{\mathrm{nsk}}), v^+ \rangle)^{1/2}$.
    if $\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \alpha(\tau, \varepsilon)(\widetilde{\lambda}_f(x_t))^{1+\tau}$ **then**
      Set $x_{t+1} = x_{\mathrm{nsk}}$, $m_{t+1} = m_t$, $v_{\mathrm{nsk}} = v^+$ and go to step 4.
    else
      Set $x_{t+1} = x_t$ and $m_{t+1} = 2m_t$.
    end if
  end if
end for

4.1 Time and memory space complexity guarantees

For conciseness, we present a succinct version of our complexity guarantees for the adaptive Newton sketch (only for the linear rate $\tau = 0$ and for the quadratic rate $\tau = 1$). A more general statement for any $\tau \in [0, 1]$ can be found in the proof of Theorem 3.

Theorem 3 (Geometric convergence guarantees of the adaptive Newton sketch)**.** *Let $\tau \in [0, 1]$, $p_0 \in (0, 1/2)$ and $\delta \in (0, 1/2)$. Let $\overline{m}_0$ be an initial sketch size. Then, it holds with probability at least $1 - p_0$ that Algorithm 2 returns a δ-approximate solution $\widetilde{x}$ in function value (i.e., $f(\widetilde{x}) - f(x^*) \leqslant \delta$) in less than $\overline{T} = \mathcal{O}\Big(T_{\tau, \alpha(\tau,\varepsilon), \frac{\delta}{d}} \log(\overline{d}_\mu)\Big)$ iterations, with final sketch size bounded by $2 \cdot \overline{m}$ and with total time complexity $\overline{\mathcal{C}}$. The values of $\overline{m}_0$, $\overline{m}$ and $\overline{\mathcal{C}}$ depend on the choice S as follows.*

(SRHT). *For $\tau = 1$ (quadratic rate), picking $p_0 \asymp \frac{\delta}{d}$ and assuming n large enough such that $n \gtrsim \frac{d^2 \overline{d}_\mu^2}{\delta^2}$, we have $\overline{m}_0 \asymp \frac{d}{\delta} \log(\frac{d}{\delta})$, $\overline{m} \asymp \frac{d}{\delta}(\overline{d}_\mu + \log(\frac{d}{\delta}) \log(\overline{d}_\mu))$, $\overline{T} = \mathcal{O}(\log(\overline{d}_\mu) \log\log(d/\delta))$ and*

$$\overline{\mathcal{C}} = \mathcal{O}\Big(nd \log(\overline{d}_\mu) \log(d/\delta) \log\log(d/\delta)\Big). \tag{39}$$

For $\tau = 0$ (linear rate), picking $p_0 \asymp 1/\overline{d}_\mu$ and assuming n large enough such that $n \gtrsim \frac{\overline{d}_\mu^2}{\log(\overline{d}_\mu)}$, we have $\overline{m}_0 \asymp \log^2(d/\delta)$, $\overline{m} \asymp \overline{d}_\mu$, $\overline{T} = \mathcal{O}(\log(\overline{d}_\mu) \log(d/\delta))$ and

$$\overline{\mathcal{C}} = \mathcal{O}\Big(nd \log^2(\overline{d}_\mu) \log(d/\delta)\Big). \tag{40}$$

(SJLT). *For $\tau = 1$ (quadratic rate), assuming n large enough such that $n \gtrsim \frac{\overline{d}_\mu^4 d^2 \log(\log(d/\delta))^2}{\delta^2 p_0^2}$, we have*

$\overline{m}_0 \asymp \frac{d\log(\log(d/\delta))}{p_0\delta}$, $\overline{m} \asymp \frac{d\overline{d}_\mu^2\log(\log(d/\delta))}{p_0\delta}$, $\overline{T} = \mathcal{O}(\log(\overline{d}_\mu)\log\log(d/\delta))$ *and*

$$\overline{\mathcal{C}} = \mathcal{O}\big(nd\cdot\log(\overline{d}_\mu)\cdot\log(\log(d/\delta)))\big)\ . \tag{41}$$

For $\tau = 0$ *(linear rate), assuming* n *large enough such that* $n \gtrsim \frac{\overline{d}_\mu^4\log(d/\delta)^2}{p_0^2}$, *we have* $\overline{m}_0 \asymp \frac{\log(d/\delta)}{p_0}$, $\overline{m} \asymp \frac{\overline{d}_\mu^2\log(d/\delta)}{p_0}$, $\overline{T} = \mathcal{O}(\log(\overline{d}_\mu)\log(d/\delta))$ *and*

$$\overline{\mathcal{C}} = \mathcal{O}\big(nd\cdot\log(\overline{d}_\mu)\cdot\log(d/\delta))\big)\ . \tag{42}$$

Note that adaptivity with convergence rate parameter τ comes at the cost of an additional d^τ factor for the final sketch size, compared to Algorithm 1. This is essentially due to our exit condition threshold δ/d that we choose for the following reason. For small $m \gtrsim 1$, the approximate Newton decrement $\widetilde{\lambda}_f^2(x)$ may fluctuate around $\lambda_f^2(x)$ by a factor up to $\overline{d}_\mu$ (see Theorem 1 in [13]). In this case, the exit condition $\widetilde{\lambda}_f(x_t)^2 \approx \delta$ would result in $f(x_t) - f(x^*) \approx \delta\cdot\overline{d}_\mu$. To guarantee δ-accuracy, it is sufficient to use the termination condition $\delta/\overline{d}_\mu$ to account for these fluctuations. As $\overline{d}_\mu$ is unknown, we choose to divide instead by d.

We summarize our complexity guarantees in Table 1. In contrast to gradient descent (GD), Nesterov's accelerated gradient descent (NAG) and Newton's method (NE), our time complexity has no condition number dependency and scales linearly in nd up to log-factors, and so does the original Newton sketch (NS). The NS log-factor is at least $\log(d)\log(1/\delta)$ whereas our SRHT-quadratic mode adaptive method has a log-factor $\log(\overline{d}_\mu)\log(d/\delta)\log(\log(d/\delta))$. The latter is much smaller for effective dimension $\overline{d}_\mu$ small compared to d and $1/\delta$. Furthermore, in terms of memory, our algorithm starts with small m whereas NS uses a constant sketch size $m \gtrsim d$. For $\tau = 0$, our memory savings are drastic when $\overline{d}_\mu$ is small. There are downsides to our method, in comparison to NS. For $\overline{d}_\mu$ close to d, our time complexity bounds become worse than NS, by an adaptivity-cost factor $\log\overline{d}_\mu$ for both $\tau = 0$ and $\tau = 1$. For $\tau = 1$, our *worst-case* sketch size is always greater than that of NS, by a factor $\overline{d}_\mu/\delta$, which comes from enforcing quadratic convergence.

When $\log\overline{d}_\mu \gg \log\log(d/\delta)$, then our SRHT/quadratic mode adaptive method yields a better time complexity than its linear mode counterpart, but at the expense of worse memory complexity.

In the context of ridge regression, we note that the time complexity of our SRHT-linear mode adaptive method scales similarly to the complexity of the adaptive method proposed by [20] for returning a certified δ-accurate solution. Importantly, our method applies to a much broader range of optimization problems, and can achieve better complexity by tuning the convergence rate parameter τ.

We emphasize again that our guarantees hold in a worst-case sense. In practice, the sketch size can start from a small value and may remain significantly smaller than the bounds in Table 1, which we illustrate in our numerical experiments.

5 Numerical experiments

In this section, we compare adaptive Newton Sketch (NS-ada) with other optimization methods in regularized logistic regression problems as in Example 4. The compared methods include Newton Sketch (NS) with fixed sketching dimension, Newton's method (NE), gradient descent method (GD) and Nesterov's accelerated gradient descent method (NAG) [26]. For NS-ada and NS, we consider both SJLT sketching matrices and random row sampling (RSS) sketching matrices. All numerical experiments are executed on a Dell PowerEdge R840 workstation. Specifically, we use 4 cores with 192GB ram for all compared methods.

The datasets used in the numerical experiments are collected from LIBSVM[1] [12]. The datasets for multi-class classification are manually separated into two categories. For example, in MNIST dataset, we classify even

[1]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

Table 1: We compare the time complexity of different optimization methods in order to achieve a δ-accurate solution in function value, for a function with condition number κ. 'NS-effdim' (resp. 'NS-ada') refers to our Algorithm 1 (resp. our Algorithm 2); 'linear' (resp. 'quadratic') signifies the choice $\tau = 0$ (resp. $\tau = 1$). We refer to [27] for gradient descent (GD), Nesterov's accelerated gradient method (NAG) and Newton's method (NE); we refer to [31] for the Newton sketch (NS), and we refer to Theorems 2 and 3 for our algorithms. For each algorithm, we assume that the sample size n is large enough for the time complexity to scale at least linearly in the term nd.

Algorithm	Time complexity	Sketch size	Proba.	Linear scaling regime
GD	$\kappa \cdot nd \cdot \log(1/\delta)$	-	1	-
NAG	$\sqrt{\kappa} \cdot nd \cdot \log(1/\delta)$	-	1	-
NE	$nd^2 \log(\log(1/\delta))$	-	1	-
NS	$nd \log(d) \log(1/\delta)$	d	$1 - \frac{1}{d}$	$n \gtrsim \frac{d^2}{\log d}$
NS-effdim (SHRT, linear)	$nd \log(\overline{d}_\mu) \log(1/\delta)$	$\overline{d}_\mu$	$1 - \frac{1}{\overline{d}_\mu}$	$n \gtrsim \frac{d_\mu^2}{\log(d_\mu)}$
NS-effdim (SHRT, quadratic)	$nd \log(\overline{d}_\mu/\delta) \log(\log(1/\delta))$	$\frac{\overline{d}_\mu}{\delta} \log(\overline{d}_\mu/\delta)$	$1 - \frac{1}{\overline{d}_\mu}$	$n \gtrsim \frac{\overline{d}_\mu^2 \log(\frac{\overline{d}_\mu}{\delta})}{\delta^2}$
NS-ada (SRHT, linear)	$nd \log(\overline{d}_\mu)^2 \log(d/\delta)$	$\overline{d}_\mu$	$1 - \frac{1}{\overline{d}_\mu}$	$n \gtrsim \frac{\overline{d}_\mu^2}{\log(\overline{d}_\mu)}$
NS-ada (SRHT, quadratic)	$nd \log(\overline{d}_\mu) \log(\frac{d}{\delta}) \log(\log(\frac{d}{\delta}))$	$\frac{d}{\delta}\left(\overline{d}_\mu + \log(\frac{d}{\delta}) \log(\overline{d}_\mu)\right)$	$1 - \frac{1}{\overline{d}_\mu}$	$n \gtrsim \frac{d^2 \overline{d}_\mu^2}{\delta^2}$

and odd digits. For each dataset, we randomly split it into a training set and a test set with the ratio $1:1$. Several additional numerical results and experimental details are reported in Appendix A.

5.1 Regularized logistic regression

We demonstrate the performance of all compared methods on regularized logistic regression problems. The relative error is calculated by $\frac{f(x) - f_{\text{ref}} + \epsilon}{1 + f_{\text{ref}}}$. Here f_{ref} is the minimal training loss function value among all compared methods and $\epsilon = 5 \times 10^{-7}$ is a small constant.

We report the relative error and the test error with respect to the iteration number and the cpu-time in Figures 1 and 2. NS-ada-SJLT or NS-ada-RRS can achieve the best performance in the relative error with respect to the cpu-time. We can also obverse the super-linear asymptotic convergence rate of NS-ada as it gets closer to the optimum. Compared to NS, NS-ada requires less iterations and less time to converge to a solution with small relative error. Compared to methods utilizing second-order information, first-order methods like GD and NAG are not competitive to find a high-precision solution.

5.2 Regularized logistic regression with kernel matrix

We also test on regularized logistic regression with the kernel matrix. The relative error and the test error with respect to the iteration number and the cpu-time are plotted in Figure 5 to 9. NS-ada-SJLT and NS-ada-RRS demonstrate asymptotic super-linear convergence rate of the relative error as the Newton's method in terms of iteration numbers. They also achieve a rapid decrease in relative error in terms of cpu-time. First-order methods do not perform well in terms of relative error. This may come from that the kernel matrices are

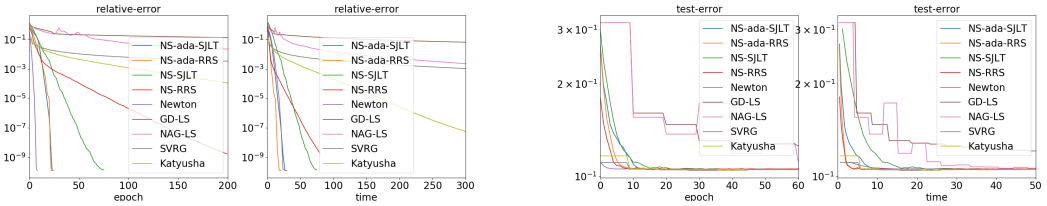

Figure 1: MNIST. $n = 30000, d = 780, \mu = 10^{-1}$.

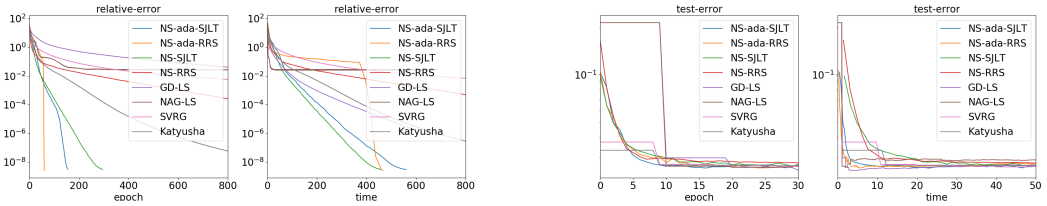

Figure 2: realsim. $n = 50000, d = 20958, \mu = 10^{-3}$.

usually ill-conditioned, i.e., with large condition number.

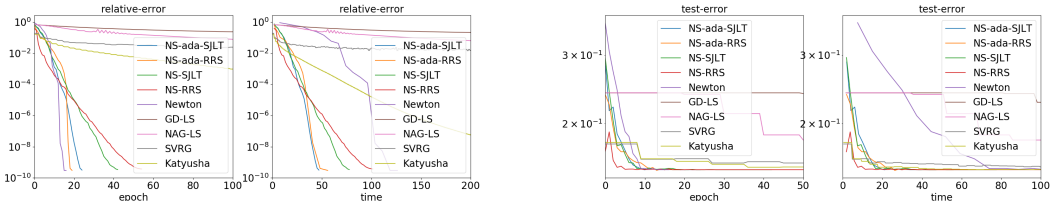

Figure 3: a8a. kernel matrix. $n = 10000, d = 10000, \mu = 10$.

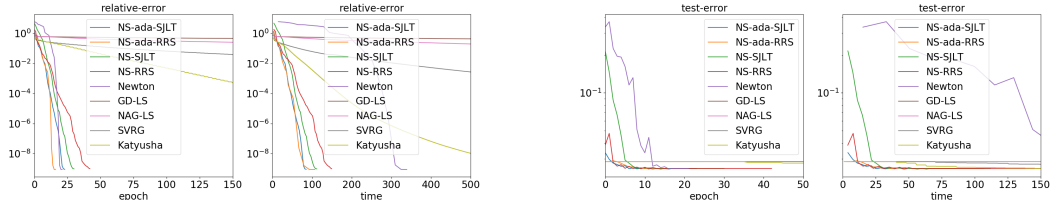

Figure 4: w7a. kernel matrix. $n = 12000, d = 12000, \mu = 10$.

References

[1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

[2] A. E. Alaoui and M. W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 775–783, 2015.

[3] H. Asi and J. C. Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019.

[4] H. Avron, K. L. Clarkson, and D. P. Woodruff. Sharper bounds for regularized data fitting. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 2017.

[5] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

[6] F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209. PMLR, 2013.

[7] B. Bartan and M. Pilanci. Distributed sketching methods for privacy preserving regression. *arXiv preprint arXiv:2002.06538*, 2020.

[8] A. S. Berahas, R. Bollapragada, and J. Nocedal. An investigation of newton-sketch and subsampled newton methods. *Optimization Methods and Software*, 35(4):661–680, 2020.

[9] R. Bollapragada, R. H. Byrd, and J. Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.

[10] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[11] R. H. Byrd, G. M. Chin, W. Neveitt, and J. Nocedal. On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995, 2011.

[12] C. Chih-Chung and L. Chih-Jen. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[13] M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal approximate matrix product in terms of stable rank. *arXiv preprint arXiv:1507.02268*, 2015.

[14] M. Derezinski, B. Bartan, M. Pilanci, and M. W. Mahoney. Debiasing distributed second order optimization with surrogate sketching and scaled regularization. In *Conference on Neural Information Processing Systems*, 2020.

[15] E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems*, pages 3670–3680, 2019.

[16] N. Doikov and P. Richtárik. Randomized block cubic Newton method. *International Conference on Machine Learning*, 2018.

[17] P. Drineas and M. W. Mahoney. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

[18] R. Gower, D. Koralev, F. Lieder, and P. Richtárik. RSN: Randomized subspace newton. In *Advances in Neural Information Processing Systems*, pages 616–625, 2019.

[19] J. Lacotte, S. Liu, E. Dobriban, and M. Pilanci. Limiting spectrum of randomized hadamard transform and optimal iterative sketching methods. In *Conference on Neural Information Processing Systems*, 2020.

[20] J. Lacotte and M. Pilanci. Effective dimension adaptive sketching methods for faster regularized least-squares optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] J. Lacotte, M. Pilanci, and M. Pavone. High-dimensional optimization in adaptive random subspaces. *arXiv preprint arXiv:1906.11809*, 2019.

[22] X. Li, S. Wang, and Z. Zhang. Do subsampled newton methods work for high-dimensional data? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4723–4730, 2020.

[23] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[24] X. Meng, M. A. Saunders, and M. W. Mahoney. Lsrn: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM Journal on Scientific Computing*, 36(2):C95–C118, 2014.

[25] J. Nelson and H. L. Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE, 2013.

[26] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[27] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[28] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

[30] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[31] M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[32] Z. Qu, P. Richtárik, M. Takác, and O. Fercoq. SDNA: Stochastic dual Newton ascent for empirical risk minimization. In *International Conference on Machine Learning*, pages 1823–1832, 2016.

[33] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

[34] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled newton methods. *Mathematical Programming*, 174(1):293–326, 2019.

[35] O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[36] S. S. Vempala. *The random projection method*, volume 65. American Mathematical Society, 2005.

[37] S. Wang, F. Roosta, P. Xu, and M. W. Mahoney. Giant: Globally improved approximate newton method for distributed optimization. *Advances in Neural Information Processing Systems*, 31:2332–2342, 2018.

[38] D. P. Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[39] P. Xu, F. Roosta, and M. W. Mahoney. Second-order optimization for non-convex machine learning: An empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2020.

[40] P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled newton methods with non-uniform sampling. *arXiv preprint arXiv:1607.00559*, 2016.

[41] Y. Yang, M. Pilanci, M. J. Wainwright, et al. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

A Additional experimental details

For NS-ada, we double the sketching dimension when $\tilde{\lambda}_f(x^{t+1}) > c_1\tilde{\lambda}_f(x^t)\min(1, c_2\tilde{\lambda}_f(x^t)^\tau)$. Here $c_1, c_2 > 0$ and $\tau \in [0, 1]$. For all compared methods, we use the backtracking line search method to find a step size satisfying the Armijo condition. For NS-ada, NS and NE, we stop the algorithm when $\tilde{\lambda}_f(x) < 10^{-6}$ or $\lambda_f(x) < 10^{-6}$. For GD and NAG, we first compute a referenced solution $\tilde{x}^*$ based on NS-ada. Then, we stop the algorithm when $\frac{f(x)-f(\tilde{x}^*)}{1+f(\tilde{x}^*)} < 10^{-6}$.

The parameters for NS-ada and NS for each dataset are summarized in Tables 2 to 4.

Dataset	m_0	c_1	τ	c_2
RCV1	100	2	0	1
MNIST	100	0.5	1	6
gisette	100	2	0	1
realsim	100	2	0	1
epsilon	100	1	0	1

Table 2: Parameters for adaptive Newton sketch with SJLT sketching.

Dataset	m_0	c_1	τ	c_2
RCV1	100	1	0	1
MNIST	100	0.5	1	6
gisette	100	2	0	1
realsim	100	2	0	1
epsilon	100	1	0	1

Table 3: Parameters for adaptive Newton sketch with RRS sketching.

Dataset	m (SJLT)	m (RRS)
RCV1	800	800
MNIST	800	1600
gisette	400	400
realsim	800	3200
epsilon	800	3200

Table 4: Sketching dimensions of Newton Sketch.

We present numerical performance of compared methods with additional details and additional numerical experiments in Figures 5 to 9. Comparatively, NS-ada-RRS tends to have larger sketching dimension than NS-ada-SJLT. This may come from that NS-RRS has stronger oscillations than NS-SJLT in the plot of $\tilde{\lambda}_f(x^t)$. Thus, NSN-ada-RRS can be slower than NS-ada-SJLT in some test cases where n is not significantly larger than d.

For kernelized regularized logistic regression, the data matrices A and $\tilde{A}$ are constructed as kernel matrices based on the original data features. Namely, it follows

$$A_{i,j} = k(\tilde{a}_i, \tilde{a}_j), \quad A^{\text{test}}_{i,j} = k(\tilde{a}^{\text{test}}_i, \tilde{a}_j),$$

where $\{\tilde{a}_i\}_{i=1}^n$ and $\{\tilde{a}_j^{\text{test}}\}_{i=1}^{n_{\text{test}}}$ are original data features from the training set and test set respectively. Here $k(x, x') : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a positive kernel function. We use the isotropic Gaussian kernel function:

$$k(x, x') = (2\pi h)^{-d/2} \exp\left(-\frac{1}{2h}\|x - x'\|_2^2\right),$$

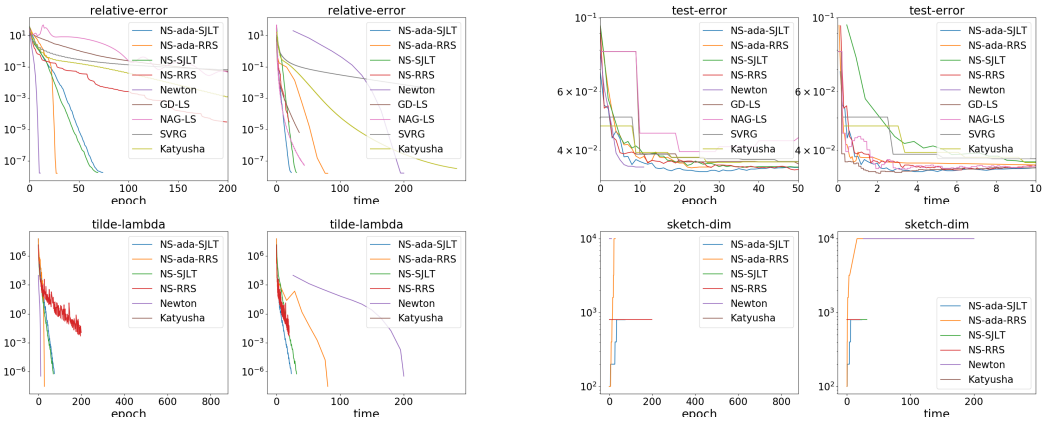

Figure 5: RCV1. $n = 10000, d = 47236, \mu = 10^{-3}$.

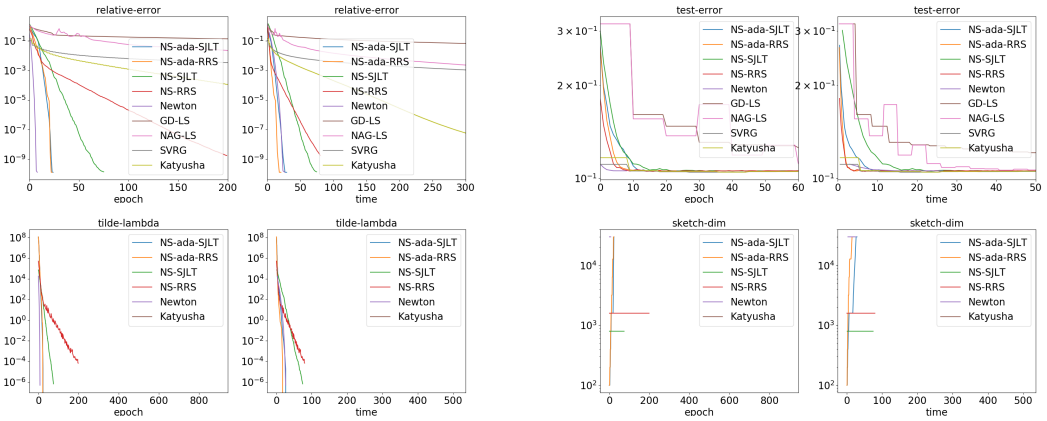

Figure 6: MNIST. $n = 30000, d = 780, \mu = 10^{-1}$.

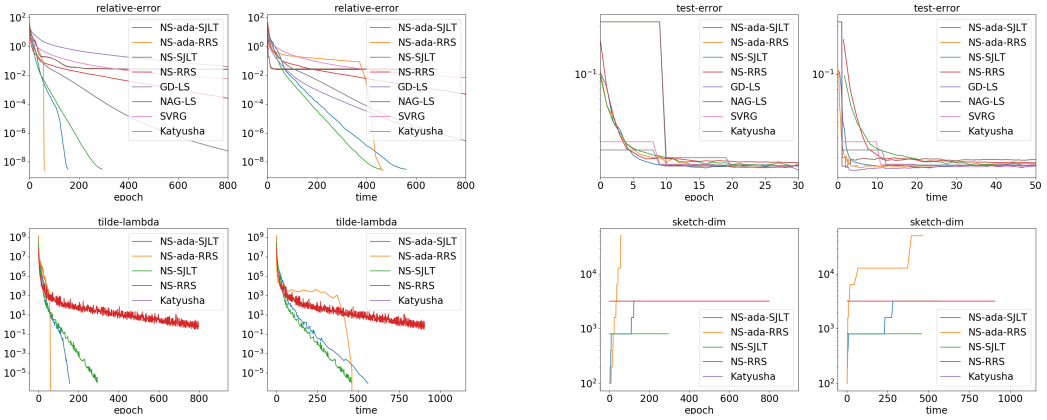

Figure 7: realsim. $n = 50000, d = 20958, \mu = 10^{-3}$.

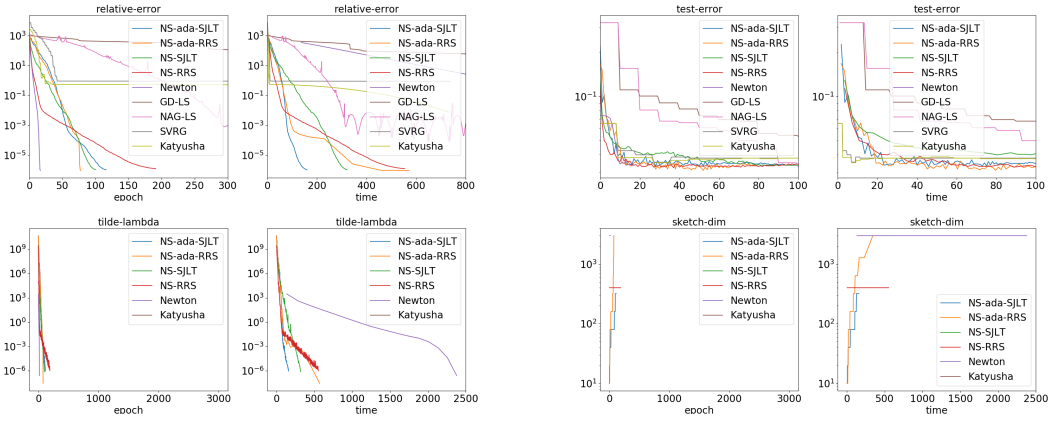

Figure 8: gisette. $n = 3000, d = 5000, \mu = 10^{-3}$.

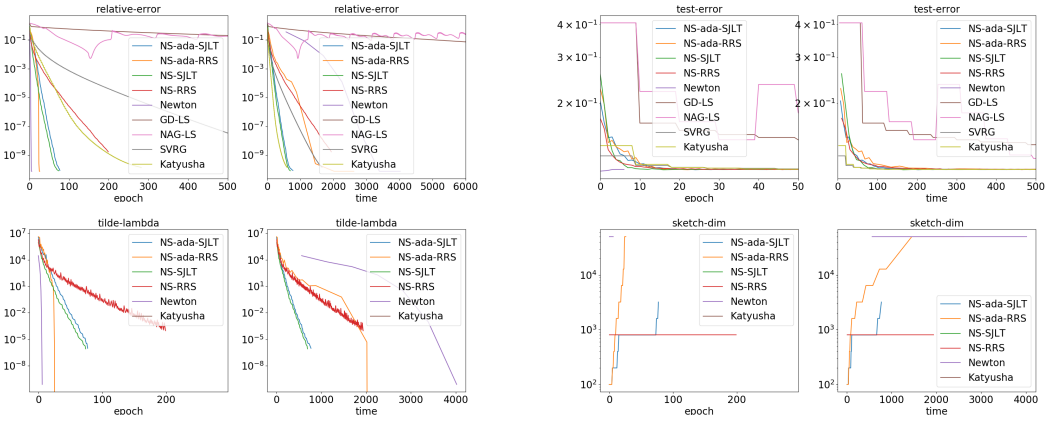

Figure 9: epsilon. $n = 50000, d = 2000, \mu = 10^{-1}$.

where $h > 0$ is the bandwidth. We set $h = 10$ for a8a dataset and $h = 20$ for w7a dataset. For NS-ada-SJLT and NS-ada-RRS, we let $c_1 = 0.5, \tau = 0$ and $c_2 = 1$. For NS, the sketching dimensions are summarized in Table 5.

Dataset	m (SJLT)	m (RRS)
a8a-kernel	100	800
w7a-kernel	100	800

Table 5: Sketching dimensions of Newton Sketch. kernel matrix.

We present numerical results with additional details in Figures 10 and 11. We can also observe super linear convergence rate of NS-ada in the plot of $\tilde{\lambda}_f(x^t)$ when x^t is close to the optimum of the optimization problem. Similarly, NS-ada-RRS tends to have larger sketching dimension than NS-ada-SJLT.

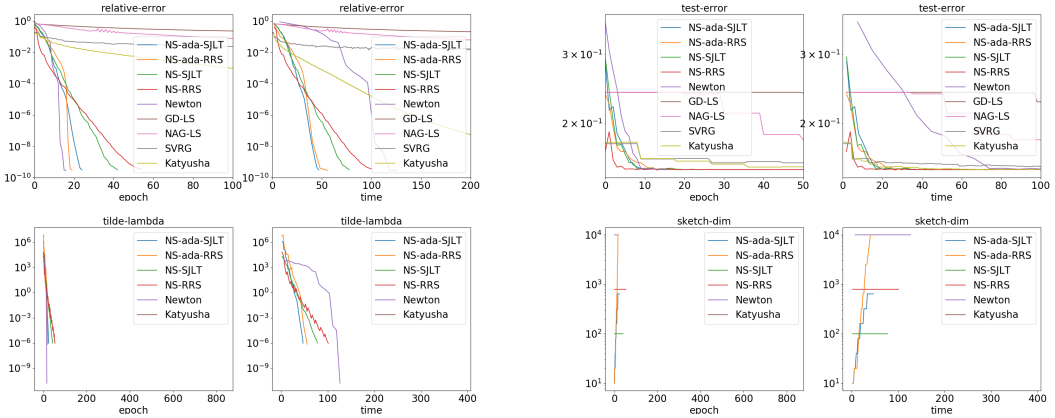

Figure 10: a8a. kernel matrix. $n = 10000, d = 10000, \mu = 10$.

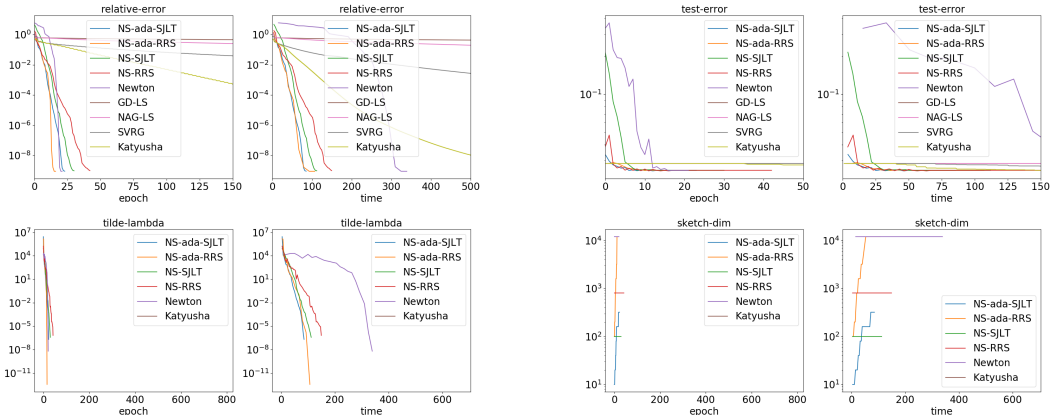

Figure 11: w7a. kernel matrix. $n = 12000, d = 12000, \mu = 10$.

B Proof of main results

B.1 Proof of Lemma 1

Let $x \in \mathbf{dom}\, f$. We use the shorthand $A := \nabla^2 f_0(x)^{1/2}$, and we let $A = U\Sigma V^\top$ be a thin SVD of A. We denote by $H^{1/2}$ an invertible square-root matrix of the Hessian $H \equiv H(x) = A^\top A + \nabla^2 g(x)$. Recall that $H_S \equiv H_S(x) = A^\top S^\top S A + \nabla^2 g(x)$. Then, we have

$$\begin{aligned} C_S = H^{-\frac{1}{2}} H_S H^{-\frac{1}{2}} &= H^{-\frac{1}{2}}(H + (H_S - H))H^{-\frac{1}{2}} \\ &= I_d + H^{-1/2}(H_S - H)H^{-1/2} \\ &= I_d + H^{-1/2} V\Sigma(U^\top S^\top S U - I_d)\Sigma V^\top H^{-1/2}\,. \end{aligned}$$

We use the shorthand $M := \Sigma V^\top H^{-1/2}$. Using the fact that $\nabla^2 g(x) \succeq \mu \cdot I_d$, it follows that

$$\|M\|_F^2 = \mathrm{trace}(\Sigma V^\top H^{-1} V \Sigma) \leqslant \mathrm{trace}(\Sigma V^\top (A^\top A + \mu I_d)^{-1} V \Sigma) = d_\mu(x)\,. \tag{43}$$

It remains to control the spectral norm of $M^\top(U^\top S^\top S U - I_d)M$.

(SJLT). It was shown in [25] that for $\varepsilon > 0$ and $p \in (0, 1/2)$, it holds with probability at least $1 - p$ that $\|M^\top(U^\top S^\top S U - I_d)M\|_2 \leqslant \varepsilon$ provided that $m \geqslant c_0 \frac{\|M\|_F^4}{\varepsilon^2 p}$, where $c_0 > 0$ is a universal constant. Note that this lower bound on the sketch size is increasing as a function of $\|M\|_F^2$. From inequality (43), it is then sufficient to have $m \geqslant c_0 \frac{d_\mu(x)^2}{\varepsilon^2 p}$ for the above inequality to hold with probability at least $1 - p$.

(SRHT). According to Theorems 1 and 9 in [13], it holds with probability at least $1-p$ that $\|M^\top(U^\top S^\top S U - I_d)M\|_2 \leqslant \varepsilon$ provided that $m \geqslant c_0 \cdot \varepsilon^{-2}\Big(\|M\|_F^2 + \log(\frac{1}{\varepsilon p})\log(\|M\|_F^2/p)\Big)$, where c_0 is a universal constant. Note that this lower bound on the sketch size is increasing as a function of $\|M\|_F^2$. From inequality (43), it is then sufficient to have $m \geqslant c_0 \cdot \varepsilon^{-2}\Big(d_\mu(x) + \log(\frac{1}{\varepsilon p})\log(d_\mu(x)/p)\Big)$ for the above inequality to hold with probability at least $1 - p$. □

B.2 Proof of Theorem 1

Let $x \in \mathbf{dom}\, f$. Plugging-in the definitions of v_{ne} and v_{nsk}, we have

$$\begin{aligned} \|v_{\mathrm{ne}} - v_{\mathrm{nsk}}\|_{H(x)} = \|H^{1/2}(v_{\mathrm{ne}} - v_{\mathrm{nsk}})\|_2 &= \|H^{1/2}(H_S^{-1}\nabla f(x) - H^{-1}\nabla f(x))\|_2 \\ &= \|(H^{1/2}H_S^{-1}H^{1/2} - I_d)H^{-1/2}\nabla f(x)\|_2 \\ &\leqslant \|C_S^{-1} - I_d\|_2 \cdot \|H^{-1/2}\nabla f(x)\|_2\,. \end{aligned}$$

Using that $\|H^{-1/2}\nabla f(x)\|_2 = \|v_{\mathrm{ne}}\|_{H(x)}$, we further obtain

$$\|v_{\mathrm{ne}} - v_{\mathrm{nsk}}\|_{H(x)} \leqslant \|C_S^{-1} - I_d\|_2 \cdot \|v_{\mathrm{ne}}\|_{H(x)}\,.$$

Under the event $\mathcal{E}_{x,m,\varepsilon}$, it holds for $\varepsilon \in (0, 1/4)$ that $(1+\varepsilon/2)^{-1} I_d \preceq C_S^{-1} \preceq (1-\varepsilon/2)^{-1} I_d$. Using the facts that $(1+\varepsilon/2)^{-1} \geqslant 1 - \varepsilon$ and $(1-\varepsilon/2)^{-1} \leqslant 1 + \varepsilon$, we obtain the inequality $\|C_S^{-1} - I_d\|_2 \leqslant \varepsilon$, whence

$$\|v_{\mathrm{ne}} - v_{\mathrm{nsk}}\|_{H(x)} \leqslant \varepsilon \cdot \|v_{\mathrm{ne}}\|_{H(x)}\,,$$

which proves the first inequality of Theorem 1. On the other hand, we have

$$\begin{aligned} \widetilde{\lambda}_f(x)^2 = \big\langle \nabla f(x), H_S^{-1}\nabla f(x)\big\rangle &= \Big\langle H^{-\frac{1}{2}}\nabla f(x), H^{\frac{1}{2}}H_S^{-1}H^{\frac{1}{2}}H^{-\frac{1}{2}}\nabla f(x)\Big\rangle \\ &= \|C_S^{-\frac{1}{2}}H^{-\frac{1}{2}}\nabla f(x)\|_2\,. \end{aligned}$$

It follows that

$$\frac{1}{\sigma_{\max}(C_S)} \cdot \lambda_f(x)^2 \leqslant \widetilde{\lambda}_f(x)^2 \leqslant \frac{1}{\sigma_{\min}(C_S)} \cdot \lambda_f(x)^2 \,.$$

Conditional on the event $\mathcal{E}_{x,m,\varepsilon}$ and using that $(1+\varepsilon/2)^{-1} \geqslant 1-\varepsilon$ and $(1-\varepsilon/2)^{-1} \leqslant 1+\varepsilon$, we obtain the claimed result, i.e.,

$$(1-\varepsilon) \cdot \lambda_f(x)^2 \leqslant \widetilde{\lambda}_f(x)^2 \leqslant (1+\varepsilon) \cdot \lambda_f(x)^2 \,.$$

□

B.3 Proof of Lemma 2

Our proof of this result closely follows the steps of the proof of Lemma 3(a) in [31]: the core arguments are the same, but we adapt the proof to our technical framework, that is, conditional on the event $\mathcal{E}_{x,m,\varepsilon}$.

The strategy of the proof is to show that the backtracking line search leads to a step size $s > 0$ such that $f(x_{\text{nsk}}) - f(x) \leqslant -\nu$. We define the univariate function $g(u) := f(x + uv_{\text{nsk}})$ and we set $\varepsilon' = \frac{2\varepsilon}{1-\varepsilon}$. We first show that $\hat{u} = \frac{1}{1+(1+\varepsilon')\widetilde{\lambda}_f(x)}$ satisfies the bound

$$g(\hat{u}) \leqslant g(0) - a\hat{u}\widetilde{\lambda}_f^2(x) \,, \tag{44}$$

which implies that $\hat{u}$ satisfies the exit condition of backtracking line search. Therefore, the step size s must be lower bounded as $s \geqslant b\hat{u}$, which further implies that the new iterate $x_{\text{nsk}} = x + sv_{\text{nsk}}$ satisfies the decrement bound

$$f(x_{\text{nsk}}) - f(x) \leqslant -ab \cdot \frac{\widetilde{\lambda}_f(x)^2}{1 + (1 + \frac{2\varepsilon}{1-\varepsilon})\widetilde{\lambda}_f(x)} \,.$$

By assumption, $\widetilde{\lambda}_f(x) > \eta$. Using the fact that the function $u \mapsto \frac{u^2}{1+(1+\frac{2\varepsilon}{1-\varepsilon}u)}$ is monotone increasing, we get that

$$f(x_{\text{nsk}}) - f(x) \leqslant -ab \cdot \frac{\eta^2}{1 + (1 + \frac{2\varepsilon}{1-\varepsilon})\eta} = \nu \,,$$

which is exactly the claimed result. It remains to prove the claims (44).

According to Lemma 4 in [31], we have for any $u \geqslant 0$ and $\gamma \geqslant 0$ that

$$g(u) \leqslant g(0) - u\widetilde{\lambda}_f(x)^2 - \gamma - \log(1-\gamma) \,, \tag{45}$$

provided that $u\|v_{\text{nsk}}\|_{H(x)} \leqslant \gamma < 1$. By assumption, the event $\mathcal{E}_{x,m,\varepsilon}$ holds true. As a consequence of Theorem 1, we have that

$$\|v_{\text{nsk}}\|_{H(x)} \leqslant (1+\varepsilon)\lambda_f(x) \leqslant \frac{1+\varepsilon}{1-\varepsilon} \cdot \widetilde{\lambda}_f(x) = (1+\varepsilon')\widetilde{\lambda}_f(x) \,.$$

It follows that $\hat{u}\|v_{\text{nsk}}\|_{H(x)} \leqslant \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x) < 1$. Plugging-in $u = \hat{u}$ and $\gamma = \hat{u}(1+\varepsilon'\widetilde{\lambda}_f(x))$ into (45), we obtain that

$$\begin{aligned} g(\hat{u}) &\leqslant g(0) - \hat{u}\widetilde{\lambda}_f(x)^2 - \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x) - \log(1 - \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x)) \\ &= g(0) - \left\{ \hat{u}(1+\varepsilon')^2\widetilde{\lambda}_f(x)^2 + \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x) + \log(1 - \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x)) - \hat{u}((1+\varepsilon')^2 - 1)\widetilde{\lambda}_f(x)^2 \right\} . \end{aligned}$$

Using that $\hat{u}(1+\varepsilon')^2\widetilde{\lambda}_f(x)^2 + \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x) = (1+\varepsilon')\widetilde{\lambda}_f(x)$ and $\hat{u}((1+\varepsilon')^2-1)\widetilde{\lambda}_f(x)^2 = \frac{(\varepsilon'^2+2\varepsilon')\widetilde{\lambda}_f(x)^2}{1+(1+\varepsilon')\widetilde{\lambda}_f(x)}$, we find that

$$g(\hat{u}) \leqslant g(0) - (1+\varepsilon')\widetilde{\lambda}_f(x) + \log(1+(1+\varepsilon')\widetilde{\lambda}_f(x)) + \frac{(\varepsilon'^2+2\varepsilon')\widetilde{\lambda}_f(x)^2}{1+(1+\varepsilon')\widetilde{\lambda}_f(x)}\,.$$

Applying the inequality $-z+\log(1+z) \leqslant -\frac{1}{2}\frac{z^2}{(1+z)}$ with $z = (1+\varepsilon')\widetilde{\lambda}_f(x)$, we further obtain that

$$\begin{aligned} g(\hat{u}) &\leqslant g(0) - \frac{\frac{1}{2}(1+\varepsilon')^2\widetilde{\lambda}_f(x)^2}{1+(1+\varepsilon')\widetilde{\lambda}_f(x)} + \frac{(\varepsilon'^2+2\varepsilon')\widetilde{\lambda}_f(x)^2}{1+(1+\varepsilon')\widetilde{\lambda}_f(x)} \\ &= g(0) - \left(\frac{1}{2} - \frac{\varepsilon'^2}{2} - \varepsilon'\right)\widetilde{\lambda}_f(x)^2\hat{u} \\ &\leqslant g(0) - a\widetilde{\lambda}_f(x)^2\hat{u}\,, \end{aligned}$$

where the final inequality follows by the assumption that $a \leqslant 1 - \frac{1}{2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2$, that is, $a \leqslant \frac{1}{2} - \frac{\varepsilon'^2}{2} - \varepsilon'$. This concludes the proof. □

B.4 Proof of Lemma 3

We recall Theorem 4.1.6 of [27] (see, also, Exercise 9.17 in [10]): it guarantees that for a step size $s > 0$ such that $|1 - s\|v_{\mathrm{nsk}}\|_{H(x)}| < 1$, we have

$$(1 - s\|v_{\mathrm{nsk}}\|_{H(x)})^2 \cdot H(x) \preceq H(x + sv_{\mathrm{nsk}}) \preceq \frac{1}{(1 - s\|v_{\mathrm{nsk}}\|_{H(x)})^2} \cdot H(x)\,. \tag{46}$$

By assumption, the event $\mathcal{E}_{x,m,\varepsilon'}$ holds. As a consequence of Theorem 1, we have $\|v_{\mathrm{nsk}}\|_{H(x)} \leqslant (1+\varepsilon')\|v_{\mathrm{ne}}\|_{H(x)}$. Plugging this bound into (46) and using that $\|v_{\mathrm{ne}}\|_{H(x)} = \lambda_f(x)$, we obtain

$$(1 - s(1+\varepsilon')\lambda_f(x))^2 \cdot H(x) \preceq H(x + sv_{\mathrm{nsk}}) \preceq \frac{1}{(1 - s(1+\varepsilon')\lambda_f(x))^2} \cdot H(x)\,, \tag{47}$$

for $s > 0$ such that $s(1+\varepsilon')\lambda_f(x) < 1$. Denote by s_{nsk} the step size obtained by backtracking line search. It satisfies $s_{\mathrm{nsk}} \leqslant 1$. Then, it holds that

$$\begin{aligned} s_{\mathrm{nsk}}(1+\varepsilon')\lambda_f(x) \leqslant (1+\varepsilon')\lambda_f(x) &\underset{(i)}{\leqslant} \frac{1+\varepsilon'}{\sqrt{1-\varepsilon'}} \cdot \widetilde{\lambda}_f(x) \\ &\underset{(ii)}{\leqslant} \frac{1+\varepsilon'}{\sqrt{1-\varepsilon'}} \cdot \eta \\ &\underset{(iii)}{<} 1\,, \end{aligned}$$

where inequality (i) follows from the assumption that $\mathcal{E}_{x,m,\varepsilon'}$ holds and from Theorem 1; inequality (ii) follows from the assumption that $\widetilde{\lambda}_f(x) \leqslant \eta$. Furthermore, we have $\varepsilon' \leqslant \varepsilon < 1/4$, as well as $\eta < 1/16$ (see Lemma 7) and this yields inequality (iii).

Using (47), we then obtain that

$$\begin{aligned}
\lambda_f(x_{\mathrm{nsk}}) &= \|H(x_{\mathrm{nsk}})^{-1/2}\nabla f(x_{\mathrm{nsk}})\|_2 \\
&\leqslant \frac{1}{(1-(1+\varepsilon')\lambda_f(x))} \cdot \|H(x)^{-1/2}\nabla f(x_{\mathrm{nsk}})\|_2 \\
&= \frac{1}{(1-(1+\varepsilon')\lambda_f(x))} \cdot \left\| H(x)^{-1/2}\left(\nabla f(x) + \int_0^1 H(x+sv_{\mathrm{nsk}})v_{\mathrm{nsk}}\mathrm{d}s\right)\right\|_2 \\
&\leqslant \frac{1}{(1-(1+\varepsilon')\lambda_f(x))} \cdot (M_1+M_2)\,,
\end{aligned}$$

where

$$\begin{aligned}
M_1 &= \left\| H(x)^{-1/2}\left(\nabla f(x) + \int_0^1 H(x+sv_{\mathrm{nsk}})v_{\mathrm{ne}}\mathrm{d}s\right)\right\|_2\,, \\
M_2 &= \left\| H(x)^{-1/2}\cdot\int_0^1 H(x+sv_{\mathrm{nsk}})(v_{\mathrm{nsk}}-v_{\mathrm{ne}})\mathrm{d}s\right\|_2\,.
\end{aligned}$$

It remains to bound the terms M_1 and M_2. Regarding M_1, we have after re-arranging and using inequality (47) that

$$\begin{aligned}
M_1 &= \left\| \int_0^1 \left(H(x)^{-1/2}H(x+sv_{\mathrm{nsk}})H(x)^{-1/2} - I_d\right)\mathrm{d}s \cdot H(x)^{1/2}v_{\mathrm{ne}}\right\|_2 \\
&\leqslant \left| \int_0^1 \frac{1}{(1-s(1+\varepsilon')\lambda_f(x))^2}\mathrm{d}s - 1\right| \cdot \left\|H(x)^{1/2}v_{\mathrm{ne}}\right\|_2 \\
&= \frac{(1+\varepsilon')\lambda_f^2(x)}{1-(1+\varepsilon')\lambda_f(x)}\,.
\end{aligned}$$

Regarding M_2, we have

$$\begin{aligned}
M_2 &= \left\| \int_0^1 H(x)^{-1/2}H(x+sv_{\mathrm{nsk}})H(x)^{-1/2}\mathrm{d}s\, H(x)^{1/2}(v_{\mathrm{nsk}}-v_{\mathrm{ne}})\right\|_2 \\
&\leqslant \left\| \int_0^1 \frac{1}{(1-s(1+\varepsilon')\lambda_f(x))^2}\mathrm{d}s \cdot H(x)^{1/2}(v_{\mathrm{nsk}}-v_{\mathrm{ne}})\right\|_2 \\
&= \frac{1}{1-(1+\varepsilon')\lambda_f(x)} \cdot \left\|H(x)^{1/2}(v_{\mathrm{nsk}}-v_{\mathrm{ne}})\right\|_2 \\
&\leqslant \frac{\varepsilon'\lambda_f(x)}{1-(1+\varepsilon')\lambda_f(x)}\,,
\end{aligned}$$

where the last inequality follows from the assumption that the event $\mathcal{E}_{x,m,\varepsilon'}$ holds and as a consequence of Theorem 1. Plugging these bounds on M_1 and M_2, we obtain that

$$\lambda_f(x_{\mathrm{nsk}}) \leqslant \frac{(1+\varepsilon')\lambda_f(x)^2 + \varepsilon'\lambda_f(x)}{(1-(1+\varepsilon')\lambda_f(x))^2}\,. \tag{48}$$

Recall that $\varepsilon' \leqslant \varepsilon\cdot\lambda_f(x)^\tau$. Combining this inequality with (48), we obtain

$$\lambda_f(x_{\mathrm{nsk}}) \leqslant \frac{(1+\varepsilon\,\lambda_f(x)^\tau)\cdot\lambda_f(x)^2 + \varepsilon\cdot\lambda_f(x)^{1+\tau}}{(1-(1+\varepsilon\,\lambda_f(x)^\tau)\lambda_f(x))^2} = \underbrace{\frac{\lambda_f(x)^{1-\tau} + \varepsilon\,\lambda_f(x) + \varepsilon}{(1-(1+\varepsilon\,\lambda_f(x)^\tau)\lambda_f(x))^2}}_{:=\alpha(\tau,x)}\cdot\lambda_f(x)^{1+\tau}\,.$$

On the event $\mathcal{E}_{x,m,\varepsilon'}$, we have according to Theorem 1 that $(1+\varepsilon)\lambda_f(x) \leqslant \frac{(1+\varepsilon)\widetilde{\lambda}_f(x)}{\sqrt{1-\varepsilon}} \leqslant \frac{(1+\varepsilon)\eta}{\sqrt{1-\varepsilon}} \leqslant \frac{1}{16}$, where the last inequality follows from Lemma 7. Hence, the denominator of $\alpha(\tau, x)$ satisfies

$$1 - (1 + \varepsilon\lambda_f(x)^\tau)\lambda_f(x) \geqslant 1 - (1+\varepsilon)\lambda_f(x) \geqslant \frac{15}{16}\,,$$

while the numerator of $\alpha(\tau, x)$ satisfies

$$\lambda_f(x)^{1-\tau} + \varepsilon\,\lambda_f(x) + \varepsilon \leqslant \frac{1}{16^{1-\tau}} + \frac{1}{32} + \frac{1}{2}$$

Combining these bounds together, we obtain that

$$\alpha(\tau, x) \leqslant \frac{8 + 1/2 + 16^\tau}{15} \leqslant 0.57 + \frac{16^\tau}{15} = \alpha(\tau)\,.$$

It is easy to verify that $\alpha(\tau)^{1/\tau} \leqslant 2$ for any $\tau \in (0, 1]$. Furthermore, for $\tau = 0$, we obtain that $\alpha(0) \approx 0.63333 \leqslant 0.64 = \frac{16}{25}$, and this concludes the proof. Note that a similar linear convergence rate was obtained for the Newton sketch provided that $m \gtrsim d$ (see Lemma 3 in [31]). □

B.5 Proof of Lemma 4

By induction, we obtain for any $t \geqslant 0$ that $\alpha^{\frac{1}{\tau}}\beta_t \leqslant (\alpha^{\frac{1}{\tau}}\eta)^{(1+\tau)^t}$. To have $\beta_t \leqslant \sqrt{\delta}$, it suffices that $(\alpha^{\frac{1}{\tau}}\eta)^{(1+\tau)^t} \leqslant \alpha^{\frac{1}{\tau}}\sqrt{\delta}$. Taking the logarithm on both sides, this yields $(1+\tau)^t \log(\alpha^{\frac{1}{\tau}}\eta) \leqslant \log(\alpha^{\frac{1}{\tau}}\sqrt{\delta})$, i.e., $(1+\tau)^t \log(1/\alpha^{\frac{1}{\tau}}\eta) \geqslant \log(1/\alpha^{\frac{1}{\tau}}\sqrt{\delta})$. By assumption, $\log(1/\alpha^{\frac{1}{\tau}}\eta) > 0$ and $\log(1/\alpha^{\frac{1}{\tau}}\sqrt{\delta}) > 0$. Therefore, after dividing both sides by $\log(1/\alpha^{\frac{1}{\tau}}\eta)$ and taking again the logarithm, we find that it is sufficient to have

$$\begin{aligned} t &\geqslant \lceil \frac{1}{\log(1+\tau)} \cdot \log\left(\frac{\log(1/\alpha^{\frac{1}{\tau}}\sqrt{\delta})}{\log(1/\alpha^{\frac{1}{\tau}}\eta)}\right) \rceil \\ &= \lceil \frac{1}{\log(1+\tau)} \cdot \log\left(\frac{1 + \frac{\tau\log(1/\delta)}{2\log(1/\alpha)}}{1 + \frac{\tau\log(1/\eta)}{\log(1/\alpha)}}\right) \rceil \\ &= T_{\tau,\alpha,\delta}\,. \end{aligned}$$

□

B.6 Proof of Theorem 2

We denote $N_1 := \frac{f(x_0) - f(x^*)}{\nu}$ and $\widetilde{p} := \frac{p_0}{\overline{T}+2}$, where $\overline{T} := N_1 + 1 + T_{\tau,\frac{3}{8}\delta}$. Recall that we pick $\varepsilon = 1/8$.

Our proof strategy proceeds as follows. In a first phase, we show that $f(x_{\text{nsk}}) - f(x) \leqslant -\nu$ until such a decrement cannot occur anymore, i.e., until $f(x_t) - f(x^*) < \nu$. Technical arguments for Phase 1 essentially follow from Lemma 2. Then, we enter a second phase where we observe a geometric decrease of the Newton decrement as described in Lemma 3.

We define

$$t := \inf\left\{k \geqslant 0 \mid \widetilde{\lambda}_f(x_k) \leqslant \eta\right\}\,,$$

According to Lemma 8, we have $t \leqslant N_1$ with probability at least $1 - N_1\widetilde{p}$.

We turn to the analysis of Phase 2. We suppose that $T_f > t$ (i.e., the algorithm has not terminated during Phase 1), we define the additional number of iterations $J := \min\{T_{\tau,\frac{3}{8}\delta}, T_f - t - 1\}$, and we introduce the event

$$\mathcal{E}^{(2)} := \left\{\mathcal{E}_{x_t,m_t,\varepsilon} \cap \bigcap_{j=0}^{J} \mathcal{E}_{x_{t+1+j},m_{t+1+j},\varepsilon\delta^{\frac{\tau}{2}}}\right\}.$$

Let us assume that $\mathcal{E}^{(2)}$ holds true, which happens with probability at least $1-(2+T_{\tau,\frac{3}{8}\delta})\widetilde{p}$ according to Corollary 1. According to Lemma 9, we have for any $j=0,\dots,J$ that $m_{t+1+j}=\overline{m}_2$ and,

$$\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1+j}) \leqslant (\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1}))^{(1+\tau)^j}\,.$$

Further, we have from Lemma 3 and Theorem 1 that $\lambda_f(x_{t+1}) \leqslant \frac{16}{25}\cdot\lambda_f(x_t) \leqslant \frac{\widetilde{\lambda}_f(x_t)}{\sqrt{1-\varepsilon}} \leqslant \frac{\eta}{\sqrt{1-\varepsilon}} \leqslant \frac{1}{16}$. Hence, $\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1}) < 1/8$. As a consequence of Lemma 4, we must have that $\lambda_f(x_{t+1+j})^2 \leqslant \frac{3}{8}\delta$ for some $j \leqslant T_{\tau,\frac{3}{8}\delta}$, which further implies that

$$\widetilde{\lambda}_f(x_{t+1+j})^2 \leqslant (1+\varepsilon)\lambda_f(x_{t+1+j})^2 \leqslant \frac{3(1+\varepsilon)}{8}\delta \leqslant \frac{3}{4}\delta\,.$$

The above inequality implies termination of the algorithm before the time $t+1+T_{\tau,\frac{3}{8}\delta}$. Using a union bound over $\{t \leqslant N_1\}$ and $\mathcal{E}^{(2)}$, we find that the algorithm terminates within $N_1+1+T_{\delta,\frac{3}{8}\delta}$ iterations with probability at least $1-(N_1+2+T_{\tau,\frac{3}{8}\delta})\widetilde{p}$.

It remains to guarantee that the algorithm returns a point $\widetilde{x}$ such that $f(\widetilde{x})-f(x^*) \leqslant \delta$. Note that the exit criterion guarantees that $\widetilde{\lambda}_f(\widetilde{x})^2 \leqslant \frac{3}{4}\delta$. Furthermore, the final sketch size $\widetilde{m}$ necessarily satisfies $\widetilde{m} \geqslant \overline{m}_1$, so that, according to Theorem 1, we have with probability at least $1-\widetilde{p}$ that $\lambda_f(\widetilde{x})^2 \leqslant \frac{1}{1-\varepsilon}\widetilde{\lambda}_f(\widetilde{x})^2 \leqslant \delta$. Self-concordance of f further implies that $f(\widetilde{x})-f(x^*) \leqslant \lambda_f(\widetilde{x})^2 \leqslant \delta$.

In conclusion, we have shown that the algorithm returns a δ-accurate solution within $N_1+1+T_{\tau,\frac{3}{8}\delta}$ iterations with probability at least $1-(N_1+3+T_{\tau,\frac{3}{8}\delta})\widetilde{p} = 1-p_0$. This concludes the proof. □

B.6.1 Complexity guarantees for the SJLT

With the SJLT, consider the quadratic convergence case, i.e., $\tau=1$. Let $p_0>0$ be a failure probability, and consider the sketch sizes

$$\overline{m}_1 \asymp \frac{\overline{d}_\mu^2 \log\log 1/\delta}{p_0}\,, \qquad \overline{m}_2 \asymp \frac{1}{\delta}\cdot\frac{\overline{d}_\mu^2 \log\log 1/\delta}{p_0}\,.$$

We observe quadratic convergence with $T_f = \mathcal{O}(\log\log(\frac{1}{\delta}))$ iterations. Further, assuming that the sketching cost $\mathcal{O}(nd)$ dominates the cost $\mathcal{O}(\overline{m}^2 d)$ of solving the randomized Newton system, i.e., $n \gtrsim \frac{\overline{d}_\mu^4 \log(\log(1/\delta))^2}{\delta^2 p_0^2}$, then the total complexity results in

$$\mathcal{C} = \mathcal{O}\big(nd\log\log 1/\delta\big)\,.$$

Similarly, we consider the linear convergence case, i.e., $\tau=0$, and pick a failure probability $p_0>0$. Consider the sketch sizes

$$\overline{m}_1 \asymp \overline{m}_2 \asymp \frac{\overline{d}_\mu^2 \log 1/\delta}{p_0}\,.$$

We observe linear convergence with $T_f = \mathcal{O}(\log\frac{1}{\delta})$ iterations. Assuming again that the sketching cost dominates the cost of solving the randomized Newton system, i.e., $n \gtrsim \frac{\overline{d}_\mu^4 \log^2(1/\delta)}{p_0^2}$, we obtain the total time complexity

$$\mathcal{C} = \mathcal{O}(nd\log(1/\delta))\,.$$

□

B.7 Proof of Lemma 5

Let $S \in \mathbb{R}^{m\times n}$ be an embedding, and $C_S := H^{-1/2} H_S H^{-1/2}$. We use the notations $A := \nabla^2 f_0(x)^{1/2}$, and we let $A = U\Sigma V^\top$ be a thin SVD of A. Then, we have

$$\begin{aligned} C_S = H^{-\frac{1}{2}} H_S H^{-\frac{1}{2}} &= H^{-\frac{1}{2}}(H + (H_S - H))H^{-\frac{1}{2}} \\ &= I_d + H^{-1/2}(H_S - H)H^{-1/2} \\ &= I_d + M^\top (U^\top S^\top S U - I_d) M \,, \end{aligned}$$

where $M := \Sigma V^\top H^{-1/2}$. According to [13], it holds that $\|M^\top (U^\top S^\top S U - I_d) M\|_2 \leqslant \frac{\overline{d}_\mu}{2}$ (i.e., $\|C_S\|_2 \leqslant 1 + \frac{\overline{d}_\mu}{2}$) with probability at least $1-p$, provided that $m \geqslant \Omega(\log^2(1/p))$ for a SRHT S, and, $m \geqslant \Omega(1/p)$ for a SJLT S.

Then, we use the fact that

$$\widetilde{\lambda}_f(x)^2 = \langle H^{-1/2}\nabla f(x), H^{1/2} H_S^{-1} H^{1/2} H^{-1/2} \nabla f(x)\rangle \geqslant \frac{1}{\|C_S\|_2} \cdot \lambda_f(x)^2 \,.$$

Conditional on $\|C_S\|_2 \leqslant 1 + \frac{\overline{d}_\mu}{2}$, it follows that

$$\lambda_f(x)^2 \leqslant \|C_S\|_2 \cdot \widetilde{\lambda}_f(x)^2 \leqslant (1 + \frac{\overline{d}_\mu}{2})\frac{\delta}{d} \leqslant \delta \,.$$

Using the self-concordance of f, we obtain that $f(x) - f(x^*) \leqslant \delta$. This concludes the proof. □

B.8 Proof of Theorem 3

We introduce the notations

$$\overline{T} = T_{\tau,\alpha(\tau,\varepsilon),\frac{\delta}{d}} + N_1 \,, \qquad \widetilde{p} = \frac{p_0}{\overline{T}} \qquad \text{and} \qquad \varepsilon' = \varepsilon \cdot \left(\frac{\delta}{(1+\varepsilon)d}\right)^{\tau/2} .$$

We consider $\overline{m}$ a sketch size such that $\mathcal{E}_{x,\overline{m},\varepsilon'}$ holds with probability at least $1-\widetilde{p}$, that is,

$$\begin{aligned} &\overline{m} = \Omega\Big(\frac{d^\tau \overline{d}_\mu^2 \overline{T}}{p_0 \delta^\tau}\Big) \qquad \text{for the SJLT} \,, \\ &\overline{m} = \Omega\bigg(\frac{d^\tau}{\delta^\tau}\Big(\overline{d}_\mu + \log\big(\frac{\overline{T} d^{\tau/2}}{p_0 \delta^{\tau/2}}\big)\log\big(\frac{\overline{d}_\mu \overline{T}}{p_0}\big)\Big)\bigg) \qquad \text{for the SRHT} \,. \end{aligned}$$

Phase 2. Let $t \geqslant 0$ be the first iteration such that $m_t \geqslant \overline{m}$, if any. Let $x \equiv x_{t+j}$ be an iterate after time t, for some $j \geqslant 0$. The sketch size is non-decreasing, whence $m \equiv m_{t+j} \geqslant \overline{m}$. We assume that $\mathcal{E}_{x,m,\varepsilon'}$ holds, and that the algorithm has not yet terminated, i.e., $\widetilde{\lambda}_f(x)^2 > \delta/d$. Note that $\varepsilon > \varepsilon'$, whence $\mathcal{E}_{x,m,\varepsilon}$ also holds. By Theorem 1, this implies in particular that $\widetilde{\lambda}_f(x)^2 \leqslant (1+\varepsilon)\lambda_f(x)^2$, and we further obtain that $\lambda_f(x)^2 > \frac{\delta}{(1+\varepsilon)d}$, i.e.,

$$\varepsilon' < \varepsilon \cdot \lambda_f(x)^\tau \,.$$

There are two possible events.

- E_1: Either $\widetilde{\lambda}_f(x) > \eta$. Using the fact that $\mathcal{E}_{x,m,\varepsilon}$ holds, it follows from Lemma 2 that $f(x_{\text{nsk}}) - f(x) \leqslant -\nu$.

- E_2: Or $\widetilde{\lambda}_f(x) \leqslant \eta$. Using the facts that $\mathcal{E}_{x,m,\varepsilon'}$ holds and that $\varepsilon' < \varepsilon\lambda_f(x)^\tau$, it follows from Lemma 3 that $\lambda_f(x_{\text{nsk}}) \leqslant \alpha(\tau) \cdot (\lambda_f(x))^{1+\tau}$. Assuming further that the event $\mathcal{E}_{x_{\text{nsk}},m,\varepsilon'}$ holds, we have according to Lemma 6 that $\widetilde{\lambda}_f(x_{\text{nsk}}) \leqslant \widetilde{\lambda}_f(x) \leqslant \eta$ and then

$$\begin{aligned}\widetilde{\lambda}_f(x_{\text{nsk}}) &\underset{(i)}{\leqslant} \sqrt{1+\varepsilon} \cdot \lambda_f(x_{\text{nsk}}) \leqslant \sqrt{1+\varepsilon} \cdot \alpha(\tau) \cdot (\lambda_f(x))^{1+\tau} \\ &\underset{(ii)}{\leqslant} \sqrt{1+\varepsilon} \cdot \alpha(\tau) \cdot (\widetilde{\lambda}_f(x)/\sqrt{1-\varepsilon})^{1+\tau} \\ &= \alpha(\tau,\varepsilon) \cdot (\widetilde{\lambda}_f(x))^{1+\tau}\,,\end{aligned}$$

  where inequalities (i) and (ii) are immediate consequences of Theorem 1.

Hence, conditional on E_2 occurs once, then the event E_2 occurs K additional times in a row with probability at least $1 - K\widetilde{p}$. According to Lemma 4, if $K \geqslant T_{\tau,\alpha(\tau,\varepsilon),\frac{\delta}{d}}$ then the algorithm terminates. On the other hand, the event E_1 can occur at most N_1 times.

In summary, conditional on $m_t \geqslant \overline{m}$, the algorithm must terminate within $\overline{T}$ additional iterations with probability at least $1 - \overline{T}\widetilde{p} = 1 - p_0$, and with final sketch size $m \leqslant 2\overline{m}$.

Phase 1. At each iteration, one of the following events must occur:

$$\begin{aligned}e_1 &:= \{\widetilde{\lambda}_f(x) > \eta,\ f(x_{\text{nsk}}) - f(x) \leqslant -\nu\} \\ e_2 &:= \{\widetilde{\lambda}_f(x) \leqslant \eta,\ \widetilde{\lambda}_f(x_{\text{nsk}}) \leqslant \alpha(\tau,\varepsilon)(\widetilde{\lambda}_f(x))^{1+\tau}\} \\ e_3 &:= \{m \leftarrow 2m\}\,.\end{aligned}$$

Fix any iteration $t \geqslant 0$, and suppose that the algorithm has not yet terminated. Consider the sequence of events $c_0, \ldots, c_t \in \{e_1, e_2, e_3\}$ up to time t. According to Lemma 4, any subsequence of $\{c_j\}_{j=0}^t$ which contains only the event e_2 would result in termination of Algorithm 2 if its length is greater or equal to $T_{\tau,\alpha(\tau,\varepsilon),\delta/d} + 1$. Consequently, any such subsequence must have length smaller or equal to $T_{\tau,\alpha(\tau,\varepsilon),\delta/d}$. Between two consecutive longest subsequences containing only e_2, either e_1 or e_3 occur. The event e_1 occurs at most N_1 times. By assumption on the choice of m_0, once e_3 has occurred at least $\mathcal{O}\big(\log(\overline{d}_\mu)\big)$ times then the sketch size is greater than $\overline{m}$. Consequently, there are at most $T_1 := \mathcal{O}\big(\big(N_1 + \log(\overline{d}_\mu)\big)T_{\tau,\alpha(\tau,\varepsilon),\delta/d}\big)$ iterations before reaching a sketch size m such that $m \geqslant \overline{m}$ without termination. In the latter case, we enter Phase 2.

Combining Phase 1 and Phase 2. Combining the two above results, we obtain with probability at least $1 - p_0$ that Algorithm 2 terminates with a final sketch size m smaller than $2\overline{m}$ and within a number of iterations T scaling as

$$T = T_1 + T_2 = \mathcal{O}\big(\big(N_1 + \log(\overline{d}_\mu)\big)T_{\tau,\alpha(\tau,\varepsilon),\delta/d}\big) = \mathcal{O}\big(\log(\overline{d}_\mu) \cdot T_{\tau,\alpha(\tau,\varepsilon),\delta/d}\big)\,,$$

where the last equality holds by treating N_1 as $\mathcal{O}(1)$.

Total complexity. The worst-case complexity per iteration is given as follows.

(1) For a SJLT S, the sketching cost is at most $\mathcal{O}(nd)$ at each iteration, and forming and solving the linear system $H_S v_{\text{nsk}} = -\nabla f(x)$ with a direct method using the Woodbury identity takes time $\mathcal{O}(\overline{m}^2 d)$. Multiplying by the number of iterations, we obtain the total time complexity

$$\overline{\mathcal{C}} = \mathcal{O}\left(\left(nd + \frac{\overline{d}_\mu^4 d^{2\tau+1} T^2_{\tau,\alpha(\tau,\varepsilon),\delta/d}}{\delta^{2\tau} p_0^2}\right) \cdot \log(\overline{d}_\mu) \cdot T_{\tau,\alpha(\tau,\varepsilon),\delta/d}\right)\,.$$

For $\tau \approx 1$, we have that $T_{\tau,\alpha(\tau,\varepsilon),\delta/d} = \mathcal{O}(\log(\log(d/\delta)))$. For $n \gtrsim \frac{\overline{d}_\mu^4 d^2 \log(\log(d/\delta))^2}{\delta^2 p_0^2}$, the memory and time complexities simplify to

$$\overline{m} = \Omega\left(\frac{d\overline{d}_\mu^2 \log(\log(d/\delta))}{p_0 \delta}\right), \qquad \overline{\mathcal{C}} = \mathcal{O}\big(nd \cdot \log(\overline{d}_\mu) \cdot \log(\log(d/\delta)))\big) .$$

For $\tau = 0$, we have $T_{\tau,\alpha(\tau,\varepsilon),\delta/d} = \mathcal{O}(\log(d/\delta))$. For $n \gtrsim \frac{\overline{d}_\mu^4 \log(d/\delta)^2}{p_0^2}$, the memory and time complexities simplify to

$$\overline{m} = \Omega\left(\frac{\overline{d}_\mu^2 \log(d/\delta)}{p_0}\right), \qquad \overline{\mathcal{C}} = \mathcal{O}\big(nd \cdot \log(\overline{d}_\mu) \cdot \log(d/\delta))\big) .$$

(2) We assume for simplicity that $\overline{d}_\mu \gtrsim \log^2(\log(d/\delta))$. For the SRHT, the sketching cost is $\mathcal{O}(nd \cdot \log \overline{m})$, whereas forming and solving the Newton linear system takes time $\mathcal{O}(\overline{m}^2 d)$. Thus, the total complexity is given by

$$\overline{\mathcal{C}} = \mathcal{O}\big(\big(nd \log \overline{m} + d \cdot \overline{m}^2\big) \log(\overline{d}_\mu) \cdot T_{\tau,\alpha(\tau,\varepsilon),\delta/d}\big) .$$

For $\tau \approx 1$, we have $T_{\tau,\alpha(\tau,\varepsilon),\delta/d} = \mathcal{O}(\log(\log(d/\delta)))$. Picking $p_0 \asymp 1/\overline{d}_\mu$, we obtain the memory complexity

$$\overline{m} \asymp \frac{d}{\delta}\left(\overline{d}_\mu + \log(d/\delta)\log(\overline{d}_\mu)\right) .$$

Consequently, $\log \overline{m} \lesssim \log(d/\delta)$ and $\overline{m}^2 \lesssim \frac{d^2}{\delta^2}(\overline{d}_\mu^2 + \log^2(d/\delta)\log^2(\overline{d}_\mu))$. Hence, provided that $n \gtrsim \frac{d^2\overline{d}_\mu^2}{\delta^2}$, we obtain

$$\overline{\mathcal{C}} = \mathcal{O}\big(nd \log(d/\delta) \log(\overline{d}_\mu) \log(\log(d/\delta))\big) .$$

For $\tau = 0$, we have $T_{\tau,\alpha(\tau,\varepsilon),\delta/d} = \mathcal{O}(\log(d/\delta))$. Picking $p_0 \asymp 1/\overline{d}_\mu$, we obtain the memory complexity

$$\overline{m} \asymp \overline{d}_\mu .$$

Consequently, $\log \overline{m} \lesssim \log(\overline{d}_\mu)$ and $\overline{m}^2 \lesssim \overline{d}_\mu^2$. Assuming that $n \gtrsim \overline{d}_\mu^2 / \log(\overline{d}_\mu)$, the total time complexity is

$$\overline{\mathcal{C}} = \mathcal{O}\Big(nd \cdot \log(\overline{d}_\mu)^2 \cdot \log(d/\delta)\Big) .$$

This concludes the proof. □

C Auxiliary results

Lemma 6. *Let $x \in \mathbf{dom}\, f$ and $\varepsilon \in (0, 1/4)$. Suppose that the event $\mathcal{E}_{x,m,\varepsilon} \cap \mathcal{E}_{x_{\mathrm{nsk}},m_{\mathrm{nsk}},\varepsilon}$ holds, and that $\widetilde{\lambda}_f(x) \leqslant \eta$. Then, we have that*

$$\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \widetilde{\lambda}_f(x) \leqslant \eta . \tag{49}$$

Proof. By assumption, the event $\mathcal{E}_{x_{\mathrm{nsk}},m_{\mathrm{nsk}},\varepsilon}$ holds. It follows from Theorem 1 that $\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \sqrt{1+\varepsilon} \cdot \lambda_f(x_{\mathrm{nsk}})$. We have by assumption that $\mathcal{E}_{x,m,\varepsilon}$ holds and that $\widetilde{\lambda}_f(x) \leqslant \eta$. As a consequence of Lemma 3, we have

$\widetilde{\lambda}_f(x) \leqslant \frac{16}{25} \cdot \lambda_f(x)$. As a consequence of Theorem 1, we have $\lambda_f(x) \leqslant \frac{1}{\sqrt{1-\varepsilon}} \cdot \widetilde{\lambda}_f(x)$. Combining these bounds together, we obtain that

$$\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \cdot \frac{16}{25} \cdot \widetilde{\lambda}_f(x)\,.$$

Finally, using that $\varepsilon \in (0, 1/4)$, we get that $\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \cdot \frac{16}{25} \leqslant 1$, whence,

$$\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \widetilde{\lambda}_f(x) \leqslant \eta\,.$$

□

Lemma 7. *For $\varepsilon \in (0,1)$, it holds that*

$$\eta \leqslant \frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{1}{16} \leqslant \frac{1}{16}\,. \tag{50}$$

Proof. Set $\gamma = \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2$. We aim to show that $\eta \cdot \sqrt{\gamma} \leqslant 1/16$. Plugging-in the definition of η and using that $a \geqslant 0$, we have $\eta \cdot \sqrt{\gamma} = \frac{1}{8} \cdot \frac{1-\frac{\gamma}{2}-a}{\gamma} \leqslant \frac{1}{8} \cdot \frac{1-\frac{\gamma}{2}}{\gamma}$. Since $\varphi(\gamma) := \frac{1}{8} \cdot \frac{1-\frac{\gamma}{2}}{\gamma}$ is monotone decreasing and since $\gamma \geqslant 1$, we obtain that $\eta \cdot \sqrt{\gamma} \leqslant \varphi(1)$, i.e., $\eta \cdot \sqrt{\gamma} \leqslant \frac{1}{16}$. □

C.1 Technical lemmas for the proof of Theorem 2

Lemma 8 (Phase 1)**.** *It holds that*

$$\boxed{t \leqslant N_1\,, \quad \textit{with probability at least } 1 - N_1\widetilde{p}\,.}$$

Proof. Let $j < t$ be any iteration before t_1. Note by construction of Algorithm 1 that $m_j = \overline{m}_1$. Assuming that the event $\mathcal{E}_{x_j,m_j,\varepsilon}$ holds true, it follows from Lemma 2 that we observe the decrement $f(x_{\mathrm{nsk}}) - f(x_j) \leqslant -\nu$. Consequently, under the event $\mathcal{E}^{(1)} := \bigcap_{j=0}^{t-1} \mathcal{E}_{x_j,m_j,\varepsilon}$, we obtain that

$$f(x^*) - f(x_0) \leqslant f(x_t) - f(x_0) = \sum_{j=0}^{t-1} f(x_{j+1}) - f(x_j) \leqslant -t \cdot \nu\,.$$

Hence, under $\mathcal{E}^{(1)}$, we must have $t \leqslant \frac{f(x_0)-f(x^*)}{\nu}$, i.e., $t \leqslant N_1$. According to Lemma 1 and the choice of $\overline{m}_1$, each event $\mathcal{E}_{x_j,m_j,\varepsilon}$ holds with probability at least $1 - \widetilde{p}$. Using a union bound, the event $\mathcal{E}^{(1)}$ holds with probability at least $1 - N_1\widetilde{p}$. □

Lemma 9 (Phase 2)**.** *Under the assumption that $\mathcal{E}^{(2)}$ holds, we have for any $j = 0, \ldots, J$ that*

$$\begin{cases} m_{t+1+j} = \overline{m}_2\,, \\ \widetilde{\lambda}_f(x_{t+1+j}) \leqslant \eta\,, \\ \alpha(\tau)^{\frac{1}{\tau}} \lambda_f(x_{t+1+j+1}) \leqslant (\alpha(\tau)^{\frac{1}{\tau}} \lambda_f(x_{t+1+j}))^{1+\tau}\,. \end{cases}$$

Proof. We prove this claim by induction. We start with $j = 0$. By definition of the time t, we have $\widetilde{\lambda}_f(x_t) \leqslant \eta$. Therefore, by construction of Algorithm 1, we have $m_{t+1} = \overline{m}_2$. From Lemma 6 and under $\mathcal{E}^{(2)}$, we get that $\widetilde{\lambda}_f(x_{t+1}) \leqslant \widetilde{\lambda}_f(x_t) \leqslant \eta$. Furthermore, before termination, we have that $\widetilde{\lambda}_f(x_{t+1})^2 > \frac{3}{4}\delta$. It follows from Theorem 1 that

$$\lambda_f(x_{t+1})^2 \geqslant \frac{1}{1+\varepsilon} \widetilde{\lambda}_f(x_{t+1})^2 > \frac{3}{4(1+\varepsilon)}\delta = \frac{2}{3}\delta\,,$$

and this implies in particular that $\varepsilon\delta^{\tau/2} \leqslant \varepsilon(\frac{3}{2})^{\tau/2}\lambda_f(x_{t+1})^\tau \leqslant 2\varepsilon\lambda_f(x_{t+1})^\tau$. Consequently, the hypotheses of Lemma 3 are verified and we have $\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+2}) \leqslant (\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1}))^{1+\tau}$.

Now, we prove the induction hypothesis for any $j = 1, \dots, J$, assuming that it holds for $j-1$. Since $\widetilde{\lambda}_f(x_{t+1+j-1}) \leqslant \eta$, it follows by construction of Algorithm 1 that $m_{t+1+j} = \overline{m}_2$. From Lemma 6 and under $\mathcal{E}^{(2)}$, we get that $\widetilde{\lambda}_f(x_{t+1+j}) \leqslant \widetilde{\lambda}_f(x_{t+1+j-1}) \leqslant \eta$. Furthermore, before termination, we have $\widetilde{\lambda}_f(x_{t+1+j})^2 > \frac{3}{4}\delta$. It follows from Theorem 1 that

$$\lambda_f(x_{t+1+j})^2 \geqslant \frac{1}{1+\varepsilon}\widetilde{\lambda}_f(x_{t+1+j})^2 > \frac{3}{4(1+\varepsilon)}\delta = \frac{2}{3}\delta\,,$$

and this implies in particular that $\varepsilon\delta^{\tau/2} \leqslant \varepsilon(\frac{3}{2})^{\tau/2}\lambda_f(x_{t+1+j})^\tau \leqslant 2\varepsilon\lambda_f(x_{t+1+j})^\tau$. Consequently, the hypotheses of Lemma 3 are verified and we have $\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1+j+1}) \leqslant (\alpha(\tau)^{\frac{1}{\tau}}\lambda_f(x_{t+1+j}))^{1+\tau}$. □

Corollary 1. *The event $\mathcal{E}^{(2)}$ holds true with probability at least $1 - (2 + T_{\tau,\frac{3}{8}\delta})\widetilde{p}$.*

Proof. Recall that $m_t = \overline{m}_1$ by definition of the time t. According to Lemma 9, if $\mathcal{E}^{(2)}$ holds true, then $m_{t+1+j} = \overline{m}_2$ for $j = 0, \dots, J$. From Lemma 1, we have that $\mathbb{P}(\mathcal{E}_{x_t,\overline{m}_1,\varepsilon}) \geqslant 1-\widetilde{p}$ and $\mathbb{P}(\mathcal{E}_{x_{t+1+j},\overline{m}_2,\varepsilon\delta^{\tau/2}}) \geqslant 1-\widetilde{p}$. We obtain by a union bound that $\mathbb{P}(\mathcal{E}^{(2)}) \geqslant 1 - (2 + T_{\tau,\frac{3}{8}\delta})\widetilde{p}$. □