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 first-order methods, 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 of NS-ada and NS with 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	10	2	0	1
realsim	100	2	0	1
epsilon	100	1	0	1
a8a-kernel	10	0.5	0	1
w7a-kernel	10	0.5	0	1

Table 2. Parameters of 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	10	2	0	1
realsim	100	2	0	1
epsilon	100	1	0	1
a8a-kernel	10	0.5	0	1
w7a-kernel	100	0.5	0	1

Table 3. Parameters of 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
a8a-kernel	100	800
w7a-kernel	100	800

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:

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

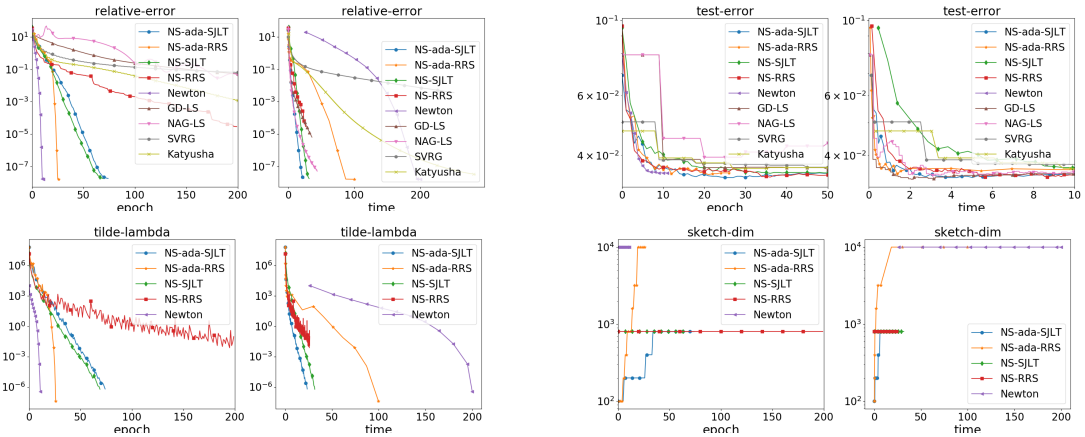

Figure 5. RCV1. $n = 10000, d = 47236, \mu = 10^{-3}$.

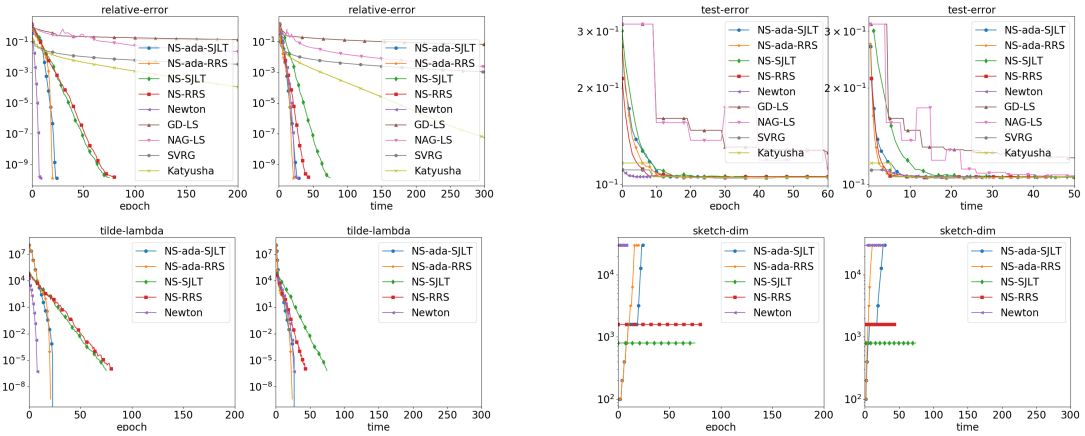

Figure 6. MNIST. $n = 30000, d = 780, \mu = 10^{-1}$.

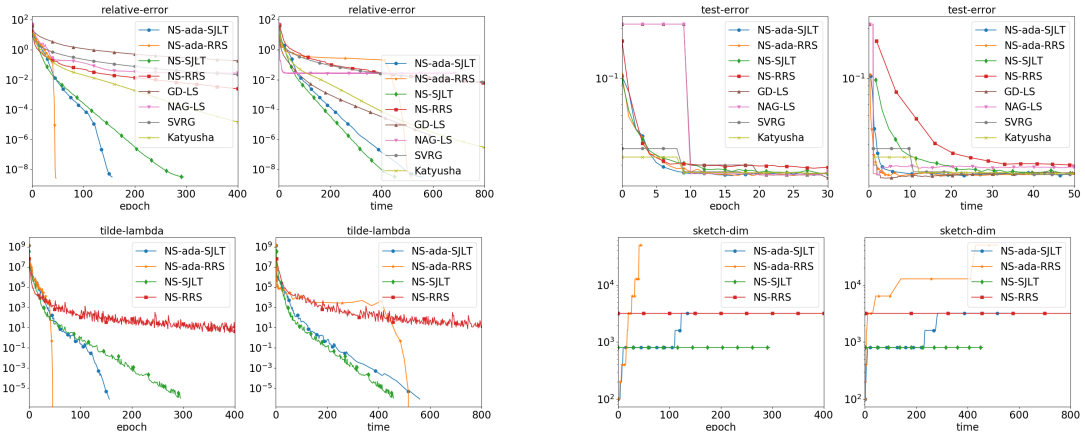

Figure 7. realsim. $n = 50000, d = 20958, \mu = 10^{-3}$.

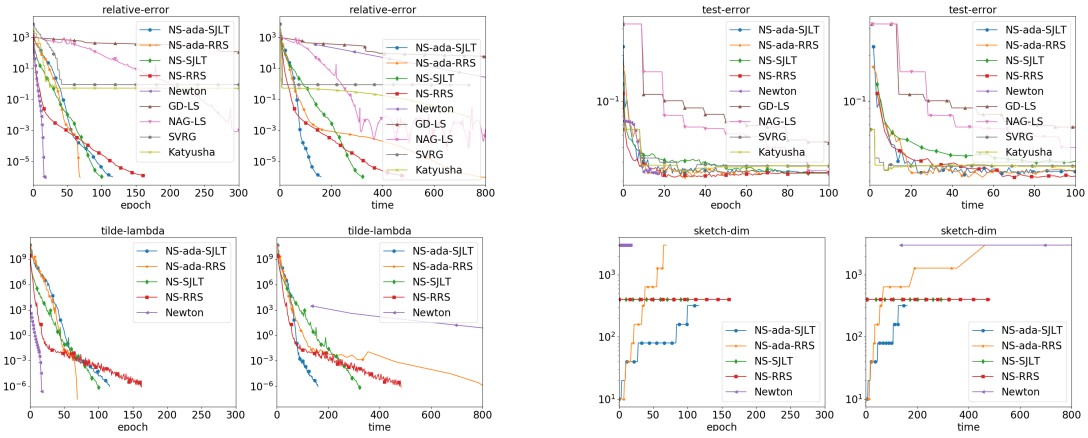

Figure 8. gisette. $n = 3000, d = 5000, \mu = 10^{-3}$.

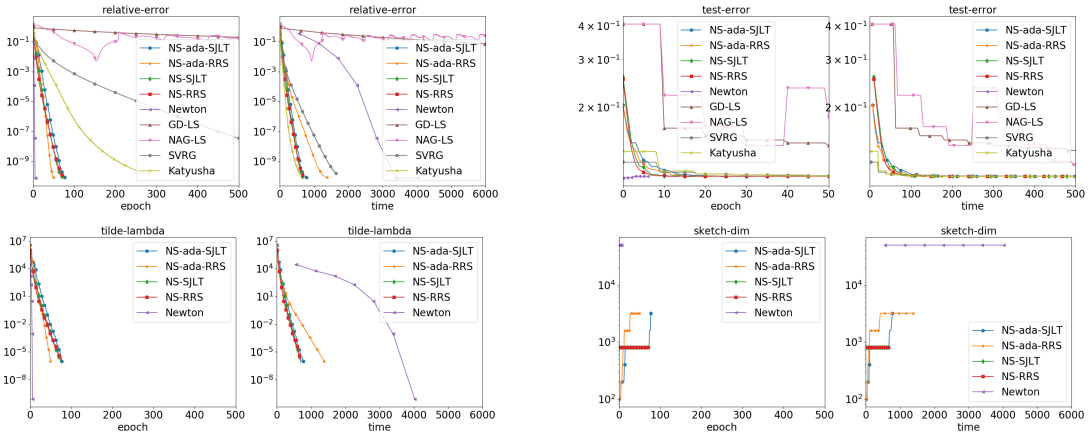

Figure 9. epsilon. $n = 50000, d = 2000, \mu = 10^{-1}$.

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.

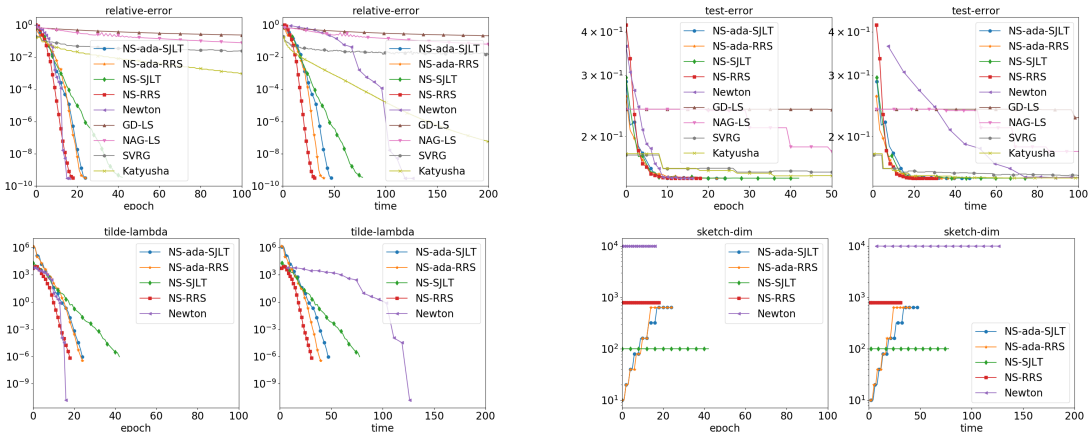

Figure 10. a8a. kernel matrix. $n = 10000, d = 10000, \mu = 10$.

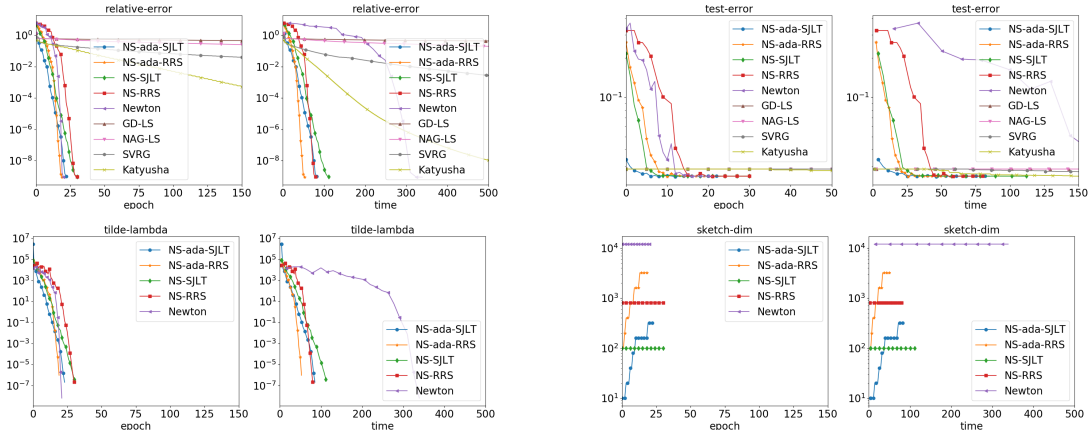

Figure 11. w7a. kernel matrix. $n = 12000, d = 12000, \mu = 10$.

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 $H_S \equiv H_S(x) =$

$A^\top S^\top SA + \nabla^2 g(x)$. Then, we have

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

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

$$\|M\|_F^2 = \text{trace}(\Sigma V^\top H^{-1} V \Sigma) \leqslant \text{trace}(\Sigma V^\top (A^\top A + \mu I_d)^{-1} V \Sigma) = d_\mu(x)\,. \tag{37}$$

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

(SJLT). It was shown in (Nelson & Nguyên, 2013) that for $\varepsilon > 0$ and $p \in (0, 1/2)$, it holds with probability at least $1 - p$ that $\|M^\top (U^\top S^\top SU - I_d) M\|_2 \leqslant \varepsilon$ 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 (37), 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 (Cohen et al., 2015), it holds with probability at least $1 - p$ that $\|M^\top (U^\top S^\top SU - I_d) M\|_2 \leqslant \varepsilon$ provided that $m \geqslant c_0\, \varepsilon^{-2} \Big(\|M\|_F^2 + \log(\frac{1}{\varepsilon p}) \log(\|M\|_F^2 / p) \Big)$, 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 (37), it is then sufficient to have $m \geqslant c_0\, \varepsilon^{-2} \Big(d_\mu(x) + \log(\frac{1}{\varepsilon p}) \log(d_\mu(x)/p) \Big)$ for the above inequality to hold with probability at least $1 - p$. □

B.2. Proof of Theorem 1

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

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

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

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

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

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

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

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

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.,

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

□

B.3. Proof of Lemma 2

Our proof of this result closely follows the steps of the proof of Lemma 3(a) in (Pilanci & Wainwright, 2017): 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 $f(x_{\mathrm{nsk}}) - f(x) \leqslant -\nu$. We define the univariate function $g(u) := f(x + u v_{\mathrm{nsk}})$ 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

$$g(\hat{u}) \leqslant g(0) - a\hat{u}\widetilde{\lambda}_f^2(x)\,, \tag{38}$$

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 $x_{\mathrm{nsk}} = x + s v_{\mathrm{nsk}}$ satisfies the decrement bound

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

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

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

which is exactly the claimed result. It remains to prove the claims (38).

According to Lemma 4 in (Pilanci & Wainwright, 2017), we have for any $u \geqslant 0$ and $\gamma \geqslant 0$ that

$$g(u) \leqslant g(0) - u\widetilde{\lambda}_f(x)^2 - \gamma - \log(1-\gamma)\,, \tag{39}$$

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

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

It follows that $\hat{u}\|v_{\mathrm{nsk}}\|_{H(x)} \leqslant \hat{u}(1+\varepsilon')\widetilde{\lambda}_f(x) < 1$. Plugging-in $u = \hat{u}$ and $\gamma = \hat{u}(1+\varepsilon'\widetilde{\lambda}_f(x))$ into (39), 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 (Nesterov, 2003) (see, also, Exercise 9.17 in (Boyd & Vandenberghe, 2004)): it guarantees that for a step size $s > 0$ such that $|1 - s\|v_{\rm nsk}\|_{H(x)}| < 1$, we have

$$(1 - s\|v_{\rm nsk}\|_{H(x)})^2 \, H(x) \preceq H(x + sv_{\rm nsk}) \preceq \frac{1}{(1 - s\|v_{\rm nsk}\|_{H(x)})^2} \, H(x) \, . \tag{40}$$

By assumption, the event $\mathcal{E}_{x,m,\varepsilon'}$ holds. As a consequence of Theorem 1, we have $\|v_{\rm nsk}\|_{H(x)} \leqslant (1 + \varepsilon')\|v_{\rm ne}\|_{H(x)}$. Plugging this bound into (40) and using that $\|v_{\rm ne}\|_{H(x)} = \lambda_f(x)$, we obtain

$$(1 - s(1 + \varepsilon')\lambda_f(x))^2 \, H(x) \preceq H(x + sv_{\rm nsk}) \preceq \frac{1}{(1 - s(1 + \varepsilon')\lambda_f(x))^2} \, H(x) \, , \tag{41}$$

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

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

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 (41), we then obtain that

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

where

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

It remains to bound the terms M_1 and M_2. Regarding M_1, we have after re-arranging and using inequality (41) that

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

Regarding M_2, we have

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

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

$$\lambda_f(x_{\mathrm{nsk}}) \leqslant \frac{(1+\varepsilon')\lambda_f(x)^2 + \varepsilon' \lambda_f(x)}{(1 - (1+\varepsilon')\lambda_f(x))^2} . \tag{42}$$

Recall that $\varepsilon' \leqslant \varepsilon \, \lambda_f(x)^\tau$. Combining this inequality with (42), we obtain

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

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 (Pilanci & Wainwright, 2017)). □

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

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

□

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 $f(x_{\mathrm{nsk}}) - f(x) \leqslant -\nu$ 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 $\lambda_f(x_{t+1}) \leqslant \frac{16}{25}\lambda_f(x_t) \leqslant \frac{\widetilde{\lambda}_f(x_t)}{\sqrt{1-\varepsilon}} \leqslant \frac{\eta}{\sqrt{1-\varepsilon}} \leqslant \frac{1}{16}$. 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

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

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 (Cohen et al., 2015), 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

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

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

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

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

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

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.,

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

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 $f(x_{\mathrm{nsk}}) - f(x) \leqslant -\nu$.
- 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 $\lambda_f(x_{\mathrm{nsk}}) \leqslant \alpha(\tau)\,(\lambda_f(x))^{1+\tau}$. Assuming further that the event $\mathcal{E}_{x_{\mathrm{nsk}},m,\varepsilon'}$ holds, we have according to Lemma 6 that $\widetilde{\lambda}_f(x_{\mathrm{nsk}}) \leqslant \widetilde{\lambda}_f(x) \leqslant \eta$ and then

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

  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:

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

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

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

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 $H_S v_{\mathrm{nsk}} = -\nabla f(x)$ 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

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

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

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

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

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

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

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

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

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

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 $\mathcal{E}_{x,m,\varepsilon} \cap \mathcal{E}_{x_{\rm nsk},m_{\rm nsk},\varepsilon}$ holds, and that $\widetilde{\lambda}_f(x) \leqslant \eta$. Then, we have that*

$$\widetilde{\lambda}_f(x_{\rm nsk}) \leqslant \widetilde{\lambda}_f(x) \leqslant \eta\,. \tag{43}$$

Proof. By assumption, the event $\mathcal{E}_{x_{\rm nsk},m_{\rm nsk},\varepsilon}$ holds. It follows from Theorem 1 that $\widetilde{\lambda}_f(x_{\rm nsk}) \leqslant \sqrt{1+\varepsilon}\,\lambda_f(x_{\rm nsk})$. 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 $\widetilde{\lambda}_f(x) \leqslant \frac{16}{25}\,\lambda_f(x)$. As a consequence of Theorem 1, we have $\lambda_f(x) \leqslant \frac{1}{\sqrt{1-\varepsilon}}\,\widetilde{\lambda}_f(x)$. Combining these bounds together, we obtain that

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

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

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

□

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

$$\eta \leqslant \frac{1-\varepsilon}{1+\varepsilon}\frac{1}{16} \leqslant \frac{1}{16}\,. \tag{44}$$

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

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

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

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 $j = 0,\dots,J$ 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 $j = 0, \ldots, J$. 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 $\mathbb{P}(\mathcal{E}^{(2)}) \geqslant 1 - (2 + T_{\tau, \frac{3}{8}\delta})\widetilde{p}$. □