Appendix

A Optimal Monteiro-Svaiter Acceleration Framework

In this section, we present some general results that hold for the Monteiro-Svaiter Acceleration framework. In particular, in the first part of this section (Section A.1), we present the proof of Proposition 1.

A.1 Proof of Proposition 1

To begin with, we establish a potential function for Algorithm 1, as shown in Proposition 2. The result is similar to Proposition 1 in [31], but for completeness we present its proof loosely following the strategy in [44, Theorem 5.3]. To simplify the notations, we use f^* to denote the optimal $f(\mathbf{x}^*)$.

Proposition 2. *Consider the iterates generated by Algorithm 1. If f is convex, then*

$$A_{k+1}(f(\mathbf{x}_{k+1})-f^*)+\frac{1}{2}\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2 \leq A_k(f(\mathbf{x}_k)-f^*)+\frac{1}{2}\|\mathbf{z}_k-\mathbf{x}^*\|^2. \tag{21}$$

Moreover, let $\sigma = \alpha_1+\alpha_2$ and we have

$$\sum_{k=0}^{N-1}\frac{a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2 \leq \frac{1}{1-\sigma^2}\|\mathbf{z}_0-\mathbf{x}^*\|^2. \tag{22}$$

Proof. Since f is convex, it holds that

$$\begin{aligned} f(\mathbf{x}_k)-f(\hat{\mathbf{x}}_{k+1})-\langle\nabla f(\hat{\mathbf{x}}_{k+1}),\mathbf{x}_k-\hat{\mathbf{x}}_{k+1}\rangle \geq 0,\\ f(\mathbf{x}^*)-f(\hat{\mathbf{x}}_{k+1})-\langle\nabla f(\hat{\mathbf{x}}_{k+1}),\mathbf{x}^*-\hat{\mathbf{x}}_{k+1}\rangle \geq 0. \end{aligned}$$

By summing up the two inequalities with weights a_k and A_k respectively, we get

$$A_k(f(\mathbf{x}_k)-f^*)-(A_k+a_k)(f(\hat{\mathbf{x}}_{k+1})-f^*)-a_k\langle\nabla f(\hat{\mathbf{x}}_{k+1}),\mathbf{x}^*-\hat{\mathbf{x}}_{k+1}-\frac{A_k}{a_k}(\hat{\mathbf{x}}_{k+1}-\mathbf{x}_k)\rangle \geq 0. \tag{23}$$

Let $\tilde{\mathbf{z}}_{k+1}=\hat{\mathbf{x}}_{k+1}+\frac{A_k}{a_k}(\hat{\mathbf{x}}_{k+1}-\mathbf{x}_k)$. By rearranging the terms, (23) can be rewritten as

$$(A_k+a_k)(f(\hat{\mathbf{x}}_{k+1})-f^*)-A_k(f(\mathbf{x}_k)-f^*) \leq a_k\langle\nabla f(\hat{\mathbf{x}}_{k+1}),\tilde{\mathbf{z}}_{k+1}-\mathbf{x}^*\rangle. \tag{24}$$

Moreover, note that the update rule for $\mathbf{z}_{k+1}$ in both (9) and (10) can be written as

$$\mathbf{z}_{k+1}-\mathbf{z}_k = -\frac{\hat{\eta}_k}{\eta_k}a_k\nabla f(\hat{\mathbf{x}}_{k+1}). \tag{25}$$

Also, since we also have $\mathbf{z}_k=\mathbf{y}_k+\frac{A_k}{a_k}(\mathbf{y}_k-\mathbf{x}_k)$ from (2), we can write

$$\begin{aligned} \tilde{\mathbf{z}}_{k+1}-\mathbf{z}_k &= \left[\hat{\mathbf{x}}_{k+1}+\frac{A_k}{a_k}(\hat{\mathbf{x}}_{k+1}-\mathbf{x}_k)\right]-\left[\mathbf{y}_k+\frac{A_k}{a_k}(\mathbf{y}_k-\mathbf{x}_k)\right]\\ &= \frac{A_k+a_k}{a_k}(\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k)=\frac{a_k}{\eta_k}(\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k), \end{aligned} \tag{26}$$

where we used the fact that $(A_k+a_k)\eta_k=a_k^2$ in the last equality (cf. (2)). Hence, combining (25) and (26) leads to

$$\|\tilde{\mathbf{z}}_{k+1}-\mathbf{z}_{k+1}\| = \|\tilde{\mathbf{z}}_{k+1}-\mathbf{z}_k-(\mathbf{z}_{k+1}-\mathbf{z}_k)\| = \frac{a_k}{\eta_k}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k+\hat{\eta}_k\nabla f(\hat{\mathbf{x}}_{k+1})\| \leq \sigma\frac{a_k}{\eta_k}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|. \tag{27}$$

where we used (8) in the last inequality. In the following, we distinguish two cases depending on $\hat{\eta}_k=\eta_k$ or $\hat{\eta}_k<\eta_k$. In both cases, we shall prove that

$$A_{k+1}(f(\mathbf{x}_{k+1})-f^*)+\frac{1}{2}\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2 \leq A_k(f(\mathbf{x}_k)-f^*)+\frac{1}{2}\|\mathbf{z}_k-\mathbf{x}^*\|^2-\frac{(1-\sigma^2)a_k^2}{2\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2. \tag{28}$$

If this is true, then Proposition 2 immediately follows. Indeed, since $\sigma < 1$, the last term in the right-hand side of (28) is negative, which implies (21). Moreover, (22) follows from summing the inequality in (28) from $k = 0$ to $N - 1$.

Case I: $\hat{\eta}_k = \eta_k$. Since by (9) we have $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_{k+1}$ and $A_{k+1} = A_k + a_k$, (24) becomes

$$A_{k+1}(f(\mathbf{x}_{k+1}) - f^*) - A_k(f(\mathbf{x}_k) - f^*) \leq a_k \langle \nabla f(\mathbf{x}_{k+1}), \tilde{\mathbf{z}}_{k+1} - \mathbf{x}^* \rangle.$$

Using $\mathbf{z}_{k+1} = \mathbf{z}_k - a_k \nabla f(\mathbf{x}_{k+1})$ in (9), we have

$$\begin{aligned}
&A_{k+1}(f(\mathbf{x}_{k+1}) - f^*) - A_k(f(\mathbf{x}_k) - f^*) \\
&\leq \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \tilde{\mathbf{z}}_{k+1} - \mathbf{x}^* \rangle \\
&= \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1} \rangle + \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \mathbf{z}_{k+1} - \mathbf{x}^* \rangle \\
&= \cancel{\frac{1}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2} + \frac{1}{2}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1}\|^2 - \frac{1}{2}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_k\|^2 \\
&\quad + \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2 - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \cancel{\frac{1}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2} \\
&\leq \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2 - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \frac{(1-\sigma^2)a_k^2}{2\eta_k^2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2,
\end{aligned} \tag{29}$$

where we used (26) and (27) in the last inequality. This immediately leads to (28) after rearranging the terms.

Case II: $\hat{\eta}_k < \eta_k$. Since $0 < \gamma_k < 1$ and $\mathbf{x}_{k+1} = \frac{(1-\gamma_k)A_k}{A_k+\gamma_k a_k}\mathbf{x}_k + \frac{\gamma_k(A_k+a_k)}{A_k+\gamma_k a_k}\hat{\mathbf{x}}_{k+1}$ according to (10), by Jensen's inequality we have $(A_k + \gamma_k a_k) f(\mathbf{x}_{k+1}) \leq \gamma_k (A_k + a_k) f(\hat{\mathbf{x}}_{k+1}) + (1-\gamma_k) A_k f(\mathbf{x}_k)$, which further implies that

$$(A_k+\gamma_k a_k)(f(\mathbf{x}_{k+1})-f^*)-A_k(f(\mathbf{x}_k)-f^*) \leq \gamma_k(A_k+a_k)(f(\hat{\mathbf{x}}_{k+1})-f^*)-\gamma_k A_k(f(\mathbf{x}_k)-f^*).$$

Moreover, since $A_{k+1} = A_k + \gamma_k a_k$ by (10), together with (24) we obtain

$$A_{k+1}(f(\mathbf{x}_{k+1}) - f^*) - A_k(f(\mathbf{x}_k) - f^*) \leq \gamma_k a_k \langle \nabla f(\hat{\mathbf{x}}_{k+1}), \tilde{\mathbf{z}}_{k+1} - \mathbf{x}^* \rangle.$$

Using $\mathbf{z}_{k+1} = \mathbf{z}_k - \gamma_k a_k \nabla f(\hat{\mathbf{x}}_{k+1})$ in (10), we follow the same reasoning as in (29) to get:

$$\begin{aligned}
&A_{k+1}(f(\mathbf{x}_{k+1}) - f^*) - A_k(f(\mathbf{x}_k) - f^*) \\
&\leq \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \tilde{\mathbf{z}}_{k+1} - \mathbf{x}^* \rangle \\
&= \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1} \rangle + \langle \mathbf{z}_k - \mathbf{z}_{k+1}, \mathbf{z}_{k+1} - \mathbf{x}^* \rangle \\
&= \cancel{\frac{1}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2} + \frac{1}{2}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1}\|^2 - \frac{1}{2}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_k\|^2 \\
&\quad + \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2 - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \cancel{\frac{1}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2} \\
&\leq \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2 - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 - \frac{(1-\sigma^2)a_k^2}{2\eta_k^2}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2,
\end{aligned}$$

which also leads to (28). □

Next, we prove a lower bound on A_N. Recall that $\mathcal{B}$ denotes the set of iteration indices where the line search scheme backtracks, i.e., $\mathcal{B} \triangleq \{k : \hat{\eta}_k < \eta_k\}$.

Lemma 3. *For any $N \geq 0$, it holds that*

$$A_N \geq \frac{1}{4}\left(\sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k}\right)^2. \tag{30}$$

Proof. To begin with, according to the update rule of A_{k+1} in (9) and (10) and the expression of a_k in (2), the sequence $\{A_k\}$ follows the dynamic:

$$A_{k+1} = \begin{cases} A_k + a_k, & \text{if } \hat{\eta}_k = \eta_k \ (k \notin \mathcal{B}); \\ A_k + \gamma_k a_k, & \text{if } \hat{\eta}_k < \eta_k \ (k \in \mathcal{B}), \end{cases} \quad \text{where } \gamma_k = \frac{\hat{\eta}_k}{\eta_k} \text{ and } a_k = \frac{\eta_k + \sqrt{\eta_k^2 + 4\eta_k A_k}}{2}.$$

Since we initialize $A_0 = 0$, we have $a_0 = \eta_0$. We further have $A_1 = \hat{\eta}_0$, since we get $A_1 = A_0 + a_0 = \hat{\eta}_0$ if $0 \notin \mathcal{B}$, while we get $A_1 = A_0 + \gamma_0 a_0 = \frac{\hat{\eta}_0}{\eta_0}\eta_0 = \hat{\eta}_0$ if $0 \in \mathcal{B}$. Moreover:

- In **Case I** where $k \notin \mathcal{B}$, we have

$$A_{k+1} = A_k + a_k = A_k + \frac{\eta_k + \sqrt{\eta_k^2 + 4\eta_k A_k}}{2} \geq A_k + \frac{\eta_k}{2} + \sqrt{\eta_k A_k} \geq \left(\sqrt{A_k} + \frac{\sqrt{\eta_k}}{2}\right)^2,$$

which further implies that $\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{\sqrt{\eta_k}}{2} = \sqrt{A_k} + \frac{\sqrt{\hat{\eta}_k}}{2}$.

- In **Case II** where $k \in \mathcal{B}$, we have $A_{k+1} = A_k + \gamma_k a_k \geq A_k$, which implies that $\sqrt{A_{k+1}} \geq \sqrt{A_k}$.

Considering the above, we obtain $\sqrt{A_N} \geq \sqrt{A_1} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \frac{\sqrt{\hat{\eta}_k}}{2}$, which leads to (30). $\square$

Lemma 3 provides a lower bound on A_N in terms of the step sizes $\hat{\eta}_k$ in those iterations where the line search scheme does not backtrack, i.e., $k \notin \mathcal{B}$. The following lemma shows how we can further prove a lower bound in terms of all the step sizes $\{\hat{\eta}_k\}_{k=0}^{N-1}$.

Lemma 4. *We have*

$$\sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\hat{\eta}_k} \leq \frac{1}{1-\sqrt{\beta}} \left(\sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k} \right). \tag{31}$$

As a corollary, we have

$$\sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k} \geq \frac{1-\sqrt{\beta}}{2-\sqrt{\beta}} \sum_{k=0}^{N-1} \sqrt{\hat{\eta}_k}. \tag{32}$$

Proof. When the line search scheme backtracks, i.e., $k \in \mathcal{B}$, we have $\hat{\eta}_k \leq \beta \eta_k$. Therefore,

$$\sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\hat{\eta}_k} \leq \sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\beta \eta_k} \leq \sum_{k=1}^{N-1} \sqrt{\beta \eta_k} = \sqrt{\beta \eta_1} + \sum_{k=1}^{N-2} \sqrt{\beta \eta_{k+1}}. \tag{33}$$

Moreover, in the update of Algorithm 1, we have $\eta_{k+1} = \hat{\eta}_k / \beta$ if $k \notin \mathcal{B}$ (cf. Line 8) and $\eta_{k+1} = \hat{\eta}_k$ otherwise (cf. Line 13). This implies that $\eta_1 \leq \hat{\eta}_0 / \beta$ and we further have

$$\begin{aligned}
\sqrt{\beta \eta_1} + \sum_{k=1}^{N-2} \sqrt{\beta \eta_{k+1}} &= \sqrt{\beta \eta_1} + \sum_{1 \leq k \leq N-2, k \notin \mathcal{B}} \sqrt{\beta \eta_{k+1}} + \sum_{1 \leq k \leq N-2, k \in \mathcal{B}} \sqrt{\beta \eta_{k+1}} \\
&\leq \sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-2, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k} + \sum_{1 \leq k \leq N-2, k \in \mathcal{B}} \sqrt{\beta \hat{\eta}_k} \\
&\leq \sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k} + \sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\beta \hat{\eta}_k}.
\end{aligned} \tag{34}$$

We combine (33) and (34) to get

$$\sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\hat{\eta}_k} \leq \sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k} + \sum_{1 \leq k \leq N-1, k \in \mathcal{B}} \sqrt{\beta \hat{\eta}_k}.$$

By rearranging the terms and simple algebraic manipulation, we obtain (31) as desired. Finally, (32) follows by adding $\sqrt{\hat{\eta}_0} + \sum_{1 \leq k \leq N-1, k \notin \mathcal{B}} \sqrt{\hat{\eta}_k}$ to both sides of (31). $\square$

Now we are ready to prove Proposition 1.

Proof of Proposition 1. By Proposition 2, the potential function $\phi_k \triangleq A_k(f(\mathbf{x}_k) - f^*) + \frac{1}{2}\|\mathbf{z}_k - \mathbf{x}^*\|^2$ is non-increasing in each iteration. Hence, via a recursive augment we have $A_N(f(\mathbf{x}_N) - f^*) \leq \phi_N \leq \cdots \leq \phi_0 = \frac{1}{2}\|\mathbf{z}_0 - \mathbf{x}^*\|^2$, which yields $f(\mathbf{x}_N) - f^* \leq \frac{\|\mathbf{z}_0 - \mathbf{x}^*\|^2}{2A_N}$. Moreover, combining Lemma 3 and (32) in Lemma 4 leads to the second inequality in Proposition 1. $\square$

A.2 Additional Supporting Lemmas

A crucial part of our analysis is to bound the path length of the sequence $\{\mathbf{y}_k\}_{k=0}^{N}$. This is done in Lemma 8. To achieve this goal we first present the results in Lemmas 5-7, which provide the required ingredients for proving the claim in Lemma 8. In our first intermediate result, we establish uniform upper bounds for the error terms $\|\mathbf{z}_k - \mathbf{x}^*\|$ and $\|\mathbf{x}_k - \mathbf{x}^*\|$.

Lemma 5. *Recall that $\sigma = \alpha_1 + \alpha_2$. For all $k \geq 0$, we have $\|\mathbf{z}_k - \mathbf{x}^*\| \leq \|\mathbf{z}_0 - \mathbf{x}^*\|$ and $\|\mathbf{x}_k - \mathbf{x}^*\| \leq \sqrt{\frac{2}{1-\sigma^2}}\|\mathbf{z}_0 - \mathbf{x}^*\|$.*

Proof. To begin with, it follows from (21) in Proposition 2 that

$$\frac{1}{2}\|\mathbf{z}_k-\mathbf{x}^*\|^2 \leq A_k(f(\mathbf{x}_k)-f^*)+\frac{1}{2}\|\mathbf{z}_k-\mathbf{x}^*\|^2 \leq A_0(f(\mathbf{x}_0)-f^*)+\frac{1}{2}\|\mathbf{z}_0-\mathbf{x}^*\|^2 = \frac{1}{2}\|\mathbf{z}_0-\mathbf{x}^*\|^2.$$

Hence, we get $\|\mathbf{z}_k - \mathbf{x}^*\| \leq \|\mathbf{z}_0 - \mathbf{x}^*\|$ for any $k \geq 0$. To show the second inequality, we distinguish two cases and in both cases we will prove that

$$A_{k+1}\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 \leq A_k\|\mathbf{x}_k-\mathbf{x}^*\|^2+(A_{k+1}-A_k)\frac{2\sigma^2 a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2+2(A_{k+1}-A_k)\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2. \tag{35}$$

Case I: $\hat{\eta}_k = \eta_k$. Recall that in the proof of Proposition 2 we defined $\tilde{\mathbf{z}}_{k+1} = \hat{\mathbf{x}}_{k+1} + \frac{A_k}{a_k}(\hat{\mathbf{x}}_{k+1} - \mathbf{x}_k)$. Since $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_{k+1}$, we have $\mathbf{x}_{k+1} = \frac{A_k}{A_k+a_k}\mathbf{x}_k + \frac{a_k}{A_k+a_k}\tilde{\mathbf{z}}_{k+1}$ and by Jensen's inequality

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq \frac{A_k}{A_k + a_k}\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \frac{a_k}{A_k + a_k}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{x}^*\|^2.$$

Furthermore, we have

$$\|\tilde{\mathbf{z}}_{k+1}-\mathbf{x}^*\|^2 \leq 2\|\tilde{\mathbf{z}}_{k+1}-\mathbf{z}_{k+1}\|^2+2\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2 \leq \frac{2\sigma^2 a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2+2\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2, \tag{36}$$

where we used (27) in the last inequality. By combining the above two inequalities, we obtain

$$(A_k + a_k)\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq A_k\|\mathbf{x}_k - \mathbf{x}^*\|^2 + a_k\frac{2\sigma^2 a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 + 2a_k\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2, \tag{37}$$

which leads to (35) (note that $A_{k+1} = A_k + a_k$ in **Case I**).

Case II: Since $\mathbf{x}_{k+1} = \frac{(1-\gamma_k)A_k}{A_k+\gamma_k a_k}\mathbf{x}_k + \frac{\gamma_k(A_k+a_k)}{A_k+\gamma_k a_k}\hat{\mathbf{x}}_{k+1}$ and $\hat{\mathbf{x}}_{k+1} = \frac{A_k}{A_k+a_k}\mathbf{x}_k + \frac{a_k}{A_k+a_k}\tilde{\mathbf{z}}_{k+1}$, we have

$$\mathbf{x}_{k+1} = \frac{A_k}{A_k + \gamma_k a_k}\mathbf{x}_k + \frac{\gamma_k a_k}{A_k + \gamma_k a_k}\tilde{\mathbf{z}}_{k+1}.$$

Similarly, by Jensen's inequality we have

$$(A_k + \gamma_k a_k)\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 \leq A_k\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \gamma_k a_k\|\tilde{\mathbf{z}}_{k+1} - \mathbf{x}^*\|^2.$$

Combining this inequality with (36), we obtain

$$(A_k+\gamma_k a_k)\|\mathbf{x}_{k+1}-\mathbf{x}^*\|^2 \leq A_k\|\mathbf{x}_k-\mathbf{x}^*\|^2+\gamma_k a_k\frac{2\sigma^2 a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2+2\gamma_k a_k\|\mathbf{z}_{k+1}-\mathbf{x}^*\|^2. \tag{38}$$

which leads to (35) (note that $A_{k+1} = A_k + \gamma_k a_k$ in **Case II**).

Now by summing (35) over $k = 0, \ldots, N-1$, we get

$$A_N\|\mathbf{x}_N - \mathbf{x}^*\|^2 \leq \sum_{k=0}^{N-1}(A_{k+1} - A_k)\frac{2\sigma^2 a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 + \sum_{k=0}^{N-1} 2(A_{k+1} - A_k)\|\mathbf{z}_{k+1} - \mathbf{x}^*\|^2 \tag{39}$$

$$\leq 2\sigma^2 \sum_{k=0}^{N-1}(A_{k+1} - A_k) \sum_{k=0}^{N-1} \frac{a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 + 2\|\mathbf{z}_0 - \mathbf{x}^*\|^2 \sum_{k=0}^{N-1}(A_{k+1} - A_k) \tag{40}$$

$$\leq \frac{2\sigma^2}{1 - \sigma^2}A_N\|\mathbf{z}_0 - \mathbf{x}^*\|^2 + 2A_N\|\mathbf{z}_0 - \mathbf{x}^*\|^2 \tag{41}$$

$$= \frac{2A_N}{1 - \sigma^2}\|\mathbf{z}_0 - \mathbf{x}^*\|^2. \tag{42}$$

Hence, this implies that $\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leq \frac{2}{1-\sigma^2}\|\mathbf{z}_0 - \mathbf{x}^*\|^2$ for any $k \geq 0$. $\square$

A key term appearing in several of our bounds is $\frac{a_{k+1}}{A_{k+1}+a_{k+1}}$. In the next lemma, we establish an upper bound for this ratio based on a factor of its previous value, for both cases of our algorithm.

Lemma 6. *Without loss of generality assume $\beta > 1/5$. In **Case I** we have $\frac{a_{k+1}}{A_{k+1}+a_{k+1}} \leq \frac{1}{\sqrt{\beta}}\frac{a_k}{A_k+a_k}$. Otherwise, in **Case II** we have $\frac{a_{k+1}}{A_{k+1}+a_{k+1}} \leq \frac{2\sqrt{\beta}}{\sqrt{\beta}+1}\frac{a_k}{A_k+a_k}$.*

Proof. By the choice of a_k in (2) we have $\eta_k(A_k + a_k) = a_k^2$ for all $k \geq 0$. As a result, we have

$$\frac{a_k}{A_k + a_k} = \frac{\eta_k}{a_k} = \frac{2\eta_k}{\eta_k + \sqrt{\eta_k^2 + 4\eta_k A_k}} = \frac{2}{1 + \sqrt{1 + 4\frac{A_k}{\eta_k}}}, \tag{43}$$

and similarly

$$\frac{a_{k+1}}{A_{k+1} + a_{k+1}} = \frac{2}{1 + \sqrt{1 + 4\frac{A_{k+1}}{\eta_{k+1}}}}. \tag{44}$$

In **Case I**, we have $\eta_{k+1} = \eta_k/\beta$ and $A_{k+1} \geq A_k$. Hence, it implies that $A_{k+1}/\eta_{k+1} \geq \beta A_k/\eta_k$, which leads to

$$\frac{a_{k+1}}{A_{k+1} + a_{k+1}} \leq \frac{2}{1 + \sqrt{1 + \frac{4\beta A_k}{\eta_k}}} \leq \frac{2}{\sqrt{\beta} + \sqrt{\beta + \frac{4\beta A_k}{\eta_k}}} = \frac{1}{\sqrt{\beta}}\frac{2}{1 + \sqrt{1 + 4\frac{A_k}{\eta_k}}} = \frac{1}{\sqrt{\beta}}\frac{a_k}{A_k + a_k}.$$

where the second inequality follows from the fact that $\beta \leq 1$.

In **Case II**, we have $\eta_{k+1} = \hat{\eta}_k = \gamma_k\eta_k$ and $A_{k+1} = A_k + \gamma_k a_k$. Since we also have $a_k \geq \eta_k$ and $\gamma_k \leq \beta$, we obtain $A_{k+1}/\eta_{k+1} \geq A_k/(\gamma_k\eta_k) + 1 \geq A_k/(\beta\eta_k) + 1$. Hence,

$$\frac{a_{k+1}}{A_{k+1} + a_{k+1}} \leq \frac{2}{1 + \sqrt{5 + \frac{4A_k}{\beta\eta_k}}} \leq \frac{2}{1 + \frac{1}{\sqrt{\beta}}\sqrt{1 + \frac{4A_k}{\eta_k}}} \leq \frac{2\sqrt{\beta}}{\sqrt{\beta}+1}\frac{2}{1 + \sqrt{1 + \frac{4A_k}{\eta_k}}} = \frac{2\sqrt{\beta}}{\sqrt{\beta}+1}\frac{a_k}{A_k + a_k},$$

where we used $\beta > 1/5$ in the second inequality and the fact that $1 + \frac{1}{\sqrt{\beta}}x \geq \frac{\sqrt{\beta}+1}{2\sqrt{\beta}}(1 + x)$ for $x \geq 1$ in the last inequality. $\square$

Next, as a corollary of Lemma 6, we establish an upper bound on the series $\sum_{k=0}^{N-1}\frac{a_k}{A_k+a_k}$. Moreover, we use this result to establish an upper bound for $\sum_{k=0}^{N-1}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|$.

Lemma 7. *We have*

$$\sum_{k=0}^{N-1}\frac{a_k}{A_k + a_k} \leq \frac{1 + 2\sqrt{\beta} - \beta}{\sqrt{\beta} - \beta}\left(1 + \log\frac{A_N}{A_1}\right). \tag{45}$$

Moreover,

$$\sum_{k=0}^{N-1}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \leq \sqrt{\frac{1}{1-\sigma^2}\frac{1 + 2\sqrt{\beta} - \beta}{\sqrt{\beta} - \beta}\left(1 + \log\frac{A_N}{A_1}\right)}\|\mathbf{z}_0 - \mathbf{x}^*\|. \tag{46}$$

Proof. Given the initial values of A_k and a_k we have

$$\sum_{k=0}^{N-1}\frac{a_k}{A_k + a_k} = 1 + \sum_{k=1}^{N-1}\frac{a_k}{A_k + a_k} = 1 + \sum_{k\in\mathcal{B},k\geq 1}\frac{a_k}{A_k + a_k} + \sum_{k\notin\mathcal{B},k\geq 1}\frac{a_k}{A_k + a_k} \tag{47}$$

Note that using the result in Lemma 6

$$\sum_{k\in\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}\leq\sum_{k=0}^{N-2}\frac{a_{k+1}}{A_{k+1}+a_{k+1}} \tag{48}$$

$$=\sum_{k\notin\mathcal{B},k\geq 0}\frac{a_{k+1}}{A_{k+1}+a_{k+1}}+\sum_{k\in\mathcal{B},k\geq 0}\frac{a_{k+1}}{A_{k+1}+a_{k+1}} \tag{49}$$

$$\leq\sum_{k\notin\mathcal{B},k\geq 0}\frac{1}{\sqrt{\beta}}\frac{a_k}{A_k+a_k}+\sum_{k\in\mathcal{B},k\geq 0}\frac{2\sqrt{\beta}}{\sqrt{\beta}+1}\frac{a_k}{A_k+a_k} \tag{50}$$

$$\leq\frac{1}{\sqrt{\beta}}+\sum_{k\notin\mathcal{B},k\geq 1}\frac{1}{\sqrt{\beta}}\frac{a_k}{A_k+a_k}+\sum_{k\in\mathcal{B},k\geq 1}\frac{2\sqrt{\beta}}{\sqrt{\beta}+1}\frac{a_k}{A_k+a_k}. \tag{51}$$

Hence, if we move the last term in the above upper bound to the left hand side and rescale both sides of the resulted inequality we obtain

$$\sum_{k\in\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}\leq\frac{1+\sqrt{\beta}}{\sqrt{\beta}-\beta}\left(1+\sum_{k\notin\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}\right). \tag{52}$$

Now, if we replace the above upper bound into (47) we obtain

$$\sum_{k=0}^{N-1}\frac{a_k}{A_k+a_k}\leq\frac{1+2\sqrt{\beta}-\beta}{\sqrt{\beta}-\beta}\left(1+\sum_{k\notin\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}\right). \tag{53}$$

Moreover, note that for $k\notin\mathcal{B}$, we have $A_{k+1}=A_k+a_k$. Hence,

$$\sum_{k\notin\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}=\sum_{k\notin\mathcal{B},k\geq 1}\left(1-\frac{A_k}{A_{k+1}}\right)\leq\sum_{k\notin\mathcal{B},k\geq 1}(\log(A_{k+1})-\log(A_k))$$
$$\leq\sum_{k=1}^{N-1}(\log(A_{k+1})-\log(A_k))=\log\frac{A_N}{A_1}.$$

Now if we replace the above upper bound, i.e., $\log\frac{A_N}{A_1}$ with $\sum_{k\notin\mathcal{B},k\geq 1}\frac{a_k}{A_k+a_k}$ into the expression in the right-hand side of (53) we obtain the result in (45).

Next, note that by Cauchy-Schwarz inequality, we have

$$\sum_{k=0}^{N-1}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|\leq\sqrt{\sum_{k=0}^{N-1}\frac{\eta_k^2}{a_k^2}\sum_{k=0}^{N-1}\frac{a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1}-\mathbf{y}_k\|^2}\leq\sqrt{\frac{1}{1-\sigma^2}\sum_{k=0}^{N-1}\frac{\eta_k^2}{a_k^2}}\|\mathbf{z}_0-\mathbf{x}^*\|,$$

where the last inequality follows from (22). Moreover, based on the expression for a_k in (2) and the result in (45) that we just proved, we have

$$\sum_{k=0}^{N-1}\frac{\eta_k^2}{a_k^2}=\sum_{k=0}^{N-1}\frac{a_k^2}{(A_k+a_k)^2}\leq\sum_{k=0}^{N-1}\frac{a_k}{A_k+a_k}\leq\frac{1+2\sqrt{\beta}-\beta}{\sqrt{\beta}-\beta}\left(1+\log\frac{A_N}{A_1}\right).$$

Combining the two inequalities above leads to (46). □

Now we are ready to present and prove Lemma 8 which characterizes a bound on the path length of the sequence $\{\mathbf{y}_k\}_{k=0}^N$

Lemma 8. *Consider the iterates generated by Algorithm 1. Then for any N,*

$$\sum_{k=0}^{N-1}\|\mathbf{y}_{k+1}-\mathbf{y}_k\|\leq C_2\left(1+\log\frac{A_N}{A_1}\right)\|\mathbf{z}_0-\mathbf{x}^*\|. \tag{54}$$

where

$$C_2=2\sqrt{\frac{1}{1-\sigma^2}\frac{1+2\sqrt{\beta}-\beta}{\sqrt{\beta}-\beta}}+\frac{1}{\sqrt{\beta}}\left(1+\sqrt{\frac{2}{1-\sigma^2}}\right)\frac{1+2\sqrt{\beta}-\beta}{\sqrt{\beta}-\beta} \tag{55}$$

Proof. By the triangle inequality, we have

$$\|\mathbf{y}_k - \mathbf{y}_{k+1}\| \leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_{k+1}\|. \tag{56}$$

We again distinguish two cases.

Case I: $\hat{\eta}_k = \eta_k$. In this case $\hat{\mathbf{x}}_{k+1} = \mathbf{x}_{k+1}$ and $\mathbf{y}_{k+1} = \frac{A_{k+1}}{A_{k+1}+a_{k+1}}\mathbf{x}_{k+1} + \frac{a_{k+1}}{A_{k+1}+a_{k+1}}\mathbf{z}_{k+1}$, hence

$$\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_{k+1}\| = \|\mathbf{x}_{k+1} - \mathbf{y}_{k+1}\| = \frac{a_{k+1}\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|}{A_{k+1} + a_{k+1}} \leq \frac{1}{\sqrt{\beta}}\left(1 + \sqrt{\frac{2}{1-\sigma^2}}\right)\frac{a_k\|\mathbf{z}_0 - \mathbf{x}^*\|}{A_k + a_k}, \tag{57}$$

where we used Lemma 6 and the fact that $\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\| \leq \|\mathbf{z}_{k+1} - \mathbf{x}^*\| + \|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq (1 + \sqrt{\frac{2}{1-\sigma^2}})\|\mathbf{z}_0 - \mathbf{x}^*\|$ in the last inequality. Therefore, using (56) and the above bound we have

$$\|\mathbf{y}_k - \mathbf{y}_{k+1}\| \leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \frac{1}{\sqrt{\beta}}\left(1 + \sqrt{\frac{2}{1-\sigma^2}}\right)\frac{a_k}{A_k + a_k}\|\mathbf{z}_0 - \mathbf{x}^*\|. \tag{58}$$

Case II: $\hat{\eta}_k < \eta_k$. Since $\mathbf{x}_{k+1} = \frac{A_k}{A_k+\gamma_k a_k}\mathbf{x}_k + \frac{\gamma_k a_k}{A_k+\gamma_k a_k}\tilde{\mathbf{z}}_{k+1}$ and $\hat{\mathbf{x}}_{k+1} = \frac{A_k}{A_k+a_k}\mathbf{x}_k + \frac{a_k}{A_k+a_k}\tilde{\mathbf{z}}_{k+1}$, we get

$$\hat{\mathbf{x}}_{k+1} = \frac{A_k}{A_k + a_k}\Big(\mathbf{x}_{k+1} + \frac{\gamma_k a_k}{A_k}(\mathbf{x}_{k+1} - \tilde{\mathbf{z}}_{k+1})\Big) + \frac{a_k}{A_k + a_k}\tilde{\mathbf{z}}_{k+1} = \frac{A_k + \gamma_k a_k}{A_k + a_k}\mathbf{x}_{k+1} + \frac{(1-\gamma_k)a_k}{A_k + a_k}\tilde{\mathbf{z}}_{k+1}.$$

Thus, given the above equality and the expression for $\mathbf{y}_{k+1}$, i.e., $\mathbf{y}_{k+1} = \frac{A_{k+1}}{A_{k+1}+a_{k+1}}\mathbf{x}_{k+1} + \frac{a_{k+1}}{A_{k+1}+a_{k+1}}\mathbf{z}_{k+1}$, we have

$$\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_{k+1}\| \leq \frac{(1-\gamma_k)a_k}{A_k + a_k}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1}\| + \left|\frac{(1-\gamma_k)a_k}{A_k + a_k} - \frac{a_{k+1}}{A_{k+1} + a_{k+1}}\right|\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|. \tag{59}$$

Moreover, based on the result in (27), we can upper bound $\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1}\|$ by $\sigma\frac{a_k}{\eta_k}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|$ which implies that

$$\frac{(1-\gamma_k)a_k}{A_k + a_k}\|\tilde{\mathbf{z}}_{k+1} - \mathbf{z}_{k+1}\| \leq \sigma\frac{(1-\gamma_k)a_k^2}{\eta_k(A_k + a_k)}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| = \sigma(1-\gamma_k)\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|$$

where the equality holds due to the definition of a_k, and the last inequality holds as both γ_k and σ are in $(0, 1)$. On the other hand, note that

$$\frac{(1-\gamma_k)a_k}{A_k + a_k} - \frac{a_{k+1}}{A_{k+1} + a_{k+1}} \leq \frac{(1-\gamma_k)a_k}{A_k + a_k} \leq \frac{a_k}{A_k + a_k}, \tag{60}$$

$$\frac{a_{k+1}}{A_{k+1} + a_{k+1}} - \frac{(1-\gamma_k)a_k}{A_k + a_k} \leq \frac{2\sqrt{\beta}a_k}{\sqrt{\beta+1}(A_k + a_k)} - \frac{(1-\gamma_k)a_k}{A_k + a_k} \leq \frac{a_k}{A_k + a_k}. \tag{61}$$

where in the second bound we used the result in Lemma 6 and the fact that $\frac{2sqrt\beta}{\sqrt{\beta+1}} < 1$. Hence, we get

$$\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_{k+1}\| \leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \frac{a_k\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\|}{A_k + a_k} \leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \left(1 + \sqrt{\frac{2}{1-\sigma^2}}\right)\frac{a_k\|\mathbf{z}_0 - \mathbf{x}^*\|}{A_k + a_k}, \tag{62}$$

where the last inequality follows from the fact $\|\mathbf{z}_{k+1} - \mathbf{x}_{k+1}\| \leq \|\mathbf{z}_{k+1} - \mathbf{x}^*\| + \|\mathbf{x}_{k+1} - \mathbf{x}^*\|$ and the bounds in Lemma 5. Now by applying the above upper bound into (56) we obtain that

$$\|\mathbf{y}_k - \mathbf{y}_{k+1}\| \leq 2\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \left(1 + \sqrt{\frac{2}{1-\sigma^2}}\right)\frac{a_k}{A_k + a_k}\|\mathbf{z}_0 - \mathbf{x}^*\|. \tag{63}$$

Considering the upper bounds established for $\|\mathbf{y}_k - \mathbf{y}_{k+1}\|$ in case I (equation (58)) and case II (equation (63)), we can conclude that

$$\|\mathbf{y}_k - \mathbf{y}_{k+1}\| \leq 2\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \frac{1}{\sqrt{\beta}}\left(1 + \sqrt{\frac{2}{1-\sigma^2}}\right)\frac{a_k}{A_k + a_k}\|\mathbf{z}_0 - \mathbf{x}^*\|. \tag{64}$$

Finally, Lemma 8 follows from summing (64) over $k = 0$ to $N - 1$ and the result of Lemma 7. $\square$

Subroutine 1 Backtracking line search

1: **Input:** iterate $\mathbf{y} \in \mathbb{R}^d$, gradient $\mathbf{g} \in \mathbb{R}^d$, Hessian approximation $\mathbf{B} \in \mathbb{S}^d_+$, initial trial step size $\eta > 0$
2: **Parameters:** line search parameters $\beta \in (0,1)$, $\alpha_1 \geq 0$ and $\alpha_2 > 0$ such that $\alpha_1 + \alpha_2 < 1$
3: Set $\hat{\eta} \leftarrow \eta$
4: Compute $\mathbf{s}_+ \leftarrow \mathsf{LinearSolver}(\mathbf{I} + \hat{\eta}\mathbf{B}, -\hat{\eta}\mathbf{g}; \alpha_1)$ and $\hat{\mathbf{x}}_+ \leftarrow \mathbf{y} + \mathbf{s}_+$
5: **while** $\|\hat{\mathbf{x}}_+ - \mathbf{y} + \hat{\eta}\nabla f(\hat{\mathbf{x}}_+)\|_2 > (\alpha_1 + \alpha_2)\|\hat{\mathbf{x}}_+ - \mathbf{y}\|_2$ **do**
6:     Set $\tilde{\mathbf{x}}_+ \leftarrow \hat{\mathbf{x}}_+$ and $\hat{\eta} \leftarrow \beta\hat{\eta}$
7:     Compute $\mathbf{s}_+ \leftarrow \mathsf{LinearSolver}(\mathbf{I} + \hat{\eta}\mathbf{B}, -\hat{\eta}\mathbf{g}; \alpha_1)$ and $\hat{\mathbf{x}}_+ \leftarrow \mathbf{y} + \mathbf{s}_+$
8: **end while**
9: **if** $\hat{\eta} = \eta$ **then**
10:     **Return** $\hat{\eta}$ and $\hat{\mathbf{x}}_+$
11: **else**
12:     **Return** $\hat{\eta}$, $\hat{\mathbf{x}}_+$ and $\tilde{\mathbf{x}}_+$
13: **end if**

B Line Search Subroutine

In this section, we provide further details on our line search subroutine in Section 3.1. For completeness, the pseudocode of our line search scheme is shown in Subroutine 1. In Section B.1, we prove that Subrountine 1 will always terminate in a finite number of steps. In Section B.2, we provide the proof of Lemma 1.

B.1 The line search subroutine terminates properly

Recall that in our line search scheme, we keep decreasing the step size $\hat{\eta}$ by a factor of β until we find a pair $(\hat{\eta}, \hat{\mathbf{x}}_+)$ satisfying (12) (also see Lines 5 and 6 in Subroutine 1). In the following lemma, we show that when the step size $\hat{\eta}$ is smaller than a certain threshold, then the pair $(\hat{\eta}, \hat{\mathbf{x}}_+)$ satisfies both conditions in (11) and (12), which further implies that Subroutine 1 will stop in a finite number of steps.

Lemma 9. *Suppose Assumption 1 holds. If $\hat{\eta} < \frac{\alpha_2}{L_1 + \|\mathbf{B}\|_{\mathrm{op}}}$ and $\hat{\mathbf{x}}_+$ is computed according to (13), then the pair $(\hat{\eta}, \hat{\mathbf{x}}_+)$ satisfies the conditions in (11) and (12).*

Proof. By Definition 1, the pair $(\hat{\eta}, \hat{\mathbf{x}}_+)$ always satisfies the condition in (11) when $\hat{\mathbf{x}}_+$ is computed from (13). Hence, in the following we only need to prove that the condition in (12) also holds. Recall that $\mathbf{g} = \nabla f(\mathbf{y})$. By Assumption 1, the function f is L_1-smooth and thus we have

$$\|\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g}\| = \|\nabla f(\hat{\mathbf{x}}_+) - \nabla f(\mathbf{y})\| \leq L_1\|\hat{\mathbf{x}}_+ - \mathbf{y}\|.$$

Moreover, by using the triangle inequality, we get

$$\|\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g} - \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y})\| \leq \|\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g}\| + \|\mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y})\| \leq (L_1 + \|\mathbf{B}\|_{\mathrm{op}})\|\hat{\mathbf{x}}_+ - \mathbf{y}\|.$$

Hence, if $\hat{\eta} \leq \frac{\alpha_2}{L_1 + \|\mathbf{B}\|_{\mathrm{op}}}$, we have

$$\hat{\eta}\|\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g} - \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y})\| \leq \alpha_2\|\hat{\mathbf{x}}_+ - \mathbf{y}\|. \tag{65}$$

Finally, by using the triangle inequality, we can combine (11) and (65) to show that

$$\begin{aligned}
\|\hat{\mathbf{x}}_+ - \mathbf{y} + \hat{\eta}\nabla f(\hat{\mathbf{x}}_+)\| &= \|\hat{\mathbf{x}}_+ - \mathbf{y} + \hat{\eta}(\mathbf{g} + \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y})) + \hat{\eta}(\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g} - \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y}))\| \\
&\leq \|\hat{\mathbf{x}}_+ - \mathbf{y} + \hat{\eta}(\mathbf{g} + \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y}))\| + \|\hat{\eta}(\nabla f(\hat{\mathbf{x}}_+) - \mathbf{g} - \mathbf{B}(\hat{\mathbf{x}}_+ - \mathbf{y}))\| \\
&\leq \alpha_1\|\hat{\mathbf{x}}_+ - \mathbf{y}\| + \alpha_2\|\hat{\mathbf{x}}_+ - \mathbf{y}\| \\
&\leq (\alpha_1 + \alpha_2)\|\hat{\mathbf{x}}_+ - \mathbf{y}\|,
\end{aligned}$$

which means the condition in (12) is satisfied. The proof is now complete. □

B.2 Proof of Lemma 1

We follow a similar proof strategy as Lemma 3 in [34]. In the first case where $k \notin \mathcal{B}$, by definition, the line search subroutine accepts the initial step size η_k, i.e., $\hat{\eta}_k = \eta_k$. In the second case where $k \in \mathcal{B}$, the line search subroutine backtracks and returns the auxiliary iterate $\tilde{\mathbf{x}}_{k+1}$, which is computed from

(13) using the step size $\tilde{\eta}_k \triangleq \hat{\eta}_k/\beta$. Since the step size $\tilde{\eta}_k$ is rejected in our line search subroutine, it implies that the pair $(\tilde{\mathbf{x}}_{k+1}, \tilde{\eta}_k)$ does not satisfy (12), i.e.,

$$\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k + \tilde{\eta}_k \nabla f(\tilde{\mathbf{x}}_{k+1})\| > (\alpha_1 + \alpha_2)\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\|. \tag{66}$$

Moreover, since we compute $\tilde{\mathbf{x}}_{k+1}$ from (13) using step size $\tilde{\eta}_k$, the pair $(\tilde{\eta}_k, \tilde{\mathbf{x}}_{k+1})$ also satisfies the condition in (11), which means

$$\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k + \tilde{\eta}_k(\nabla f(\mathbf{y}_k) + \mathbf{B}_k(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k))\| \leq \alpha_1\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\|. \tag{67}$$

Hence, by using the triangle inequality, we can combine (66) and (67) to get

$$\begin{aligned}
&\tilde{\eta}_k\|\nabla f(\tilde{\mathbf{x}}_{k+1}) - \nabla f(\mathbf{y}_k) - \mathbf{B}_k(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k)\| \\
&\geq \|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k + \tilde{\eta}_k \nabla f(\tilde{\mathbf{x}}_{k+1})\| - \|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k + \tilde{\eta}_k(\nabla f(\mathbf{y}_k) + \mathbf{B}_k(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k))\| \\
&> (\alpha_1 + \alpha_2)\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\| - \alpha_1\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \\
&= \alpha_2\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\|,
\end{aligned}$$

which implies that

$$\hat{\eta}_k = \beta\tilde{\eta}_k > \frac{\alpha_2\beta\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\|}{\|\nabla f(\tilde{\mathbf{x}}_{k+1}) - \nabla f(\mathbf{y}_k) - \mathbf{B}_k(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k)\|}.$$

This proves the first inequality in (14).

To show the second inequality in (14), first note that $\tilde{\mathbf{x}}_{k+1}$ and $\hat{\mathbf{x}}_{k+1}$ are the inexact solutions of the linear system of equations

$$(\mathbf{I} + \tilde{\eta}_k\mathbf{B}_k)(\mathbf{x} - \mathbf{y}_k) = -\tilde{\eta}_k\mathbf{g}_k \quad \text{and} \quad (\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)(\mathbf{x} - \mathbf{y}_k) = -\hat{\eta}_k\mathbf{g}_k,$$

respectively. Let $\tilde{\mathbf{x}}^*_{k+1}$ and $\hat{\mathbf{x}}^*_{k+1}$ be the exact solutions of the above linear systems, that is, $\tilde{\mathbf{x}}^*_{k+1} = \mathbf{y}_k - \tilde{\eta}_k(\mathbf{I} + \tilde{\eta}_k\mathbf{B}_k)^{-1}\mathbf{g}_k$ and $\hat{\mathbf{x}}^*_{k+1} = \mathbf{y}_k - \hat{\eta}_k(\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)^{-1}\mathbf{g}_k$. We first establish the following inequality between $\|\tilde{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\|$ and $\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\|$:

$$\|\tilde{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| \leq \frac{1}{\beta}\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\|. \tag{68}$$

This follows from

$$\|\tilde{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| = \|\tilde{\eta}_k(\mathbf{I} + \tilde{\eta}_k\mathbf{B}_k)^{-1}\mathbf{g}_k\| \leq \tilde{\eta}_k\|(\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)^{-1}\mathbf{g}_k\| = \frac{\tilde{\eta}_k}{\hat{\eta}_k}\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| = \frac{1}{\beta}\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\|,$$

where we used the fact that $(\mathbf{I} + \tilde{\eta}_k\mathbf{B}_k)^{-1} \preceq (\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)^{-1}$ in the first inequality. Furthermore, we can show that

$$(1 - \alpha_1)\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \leq \|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| \leq (1 + \alpha_1)\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|, \tag{69}$$

$$(1 - \alpha_1)\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \leq \|\tilde{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| \leq (1 + \alpha_1)\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\|. \tag{70}$$

We will only prove (69) in the following, as (70) can be proved similarly. Note that since $(\hat{\eta}_k, \hat{\mathbf{x}}_{k+1})$ satisfies the condition in (11), we can write

$$\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k + \hat{\eta}_k(\mathbf{g}_k + \mathbf{B}_k(\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k))\| = \|(\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)(\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}^*_{k+1})\| \leq \alpha_1\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|.$$

Moreover, since $\mathbf{B}_k \succeq 0$, we have $\|\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}^*_{k+1}\| \leq \|(\mathbf{I} + \hat{\eta}_k\mathbf{B}_k)(\hat{\mathbf{x}}_{k+1} - \hat{\mathbf{x}}^*_{k+1})\| \leq \alpha_1\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|$. Thus, by the triangle inequality, we obtain

$$\begin{aligned}
\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| &\leq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| + \|\hat{\mathbf{x}}^*_{k+1} - \hat{\mathbf{x}}_{k+1}\| \leq (1 + \alpha_1)\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|. \\
\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| &\geq \|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\| - \|\hat{\mathbf{x}}^*_{k+1} - \hat{\mathbf{x}}_{k+1}\| \geq (1 - \alpha_1)\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|.
\end{aligned}$$

which proves (69). Finally, by combining (68), (69) and (70), we conclude that

$$\|\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k\| \leq \frac{1}{1 - \alpha_1}\|\tilde{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| \leq \frac{1}{(1 - \alpha_1)\beta}\|\hat{\mathbf{x}}^*_{k+1} - \mathbf{y}_k\| \leq \frac{1 + \alpha_1}{(1 - \alpha_1)\beta}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|.$$

This completes the proof.

C Hessian Approximation Update

In this section, we fully describe our Hessian approximation update in Section 3.2. We first prove Lemma 2 in Section C.1.

C.1 Proof of Lemma 2

We decompose the sum $\sum_{k=0}^{N-1} \frac{1}{\hat{\eta}_k^2}$ as

$$\sum_{k=0}^{N-1} \frac{1}{\hat{\eta}_k^2} = \frac{1}{\hat{\eta}_0^2} + \sum_{1\leq k\leq N-1, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2} + \sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{1}{\hat{\eta}_k^2} \tag{71}$$

Recall that we have $\hat{\eta}_k = \eta_k$ for $k \notin \mathcal{B}$. Hence, we can further bound the last term by

$$\begin{aligned}\sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{1}{\hat{\eta}_k^2} = \sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{1}{\eta_k^2} &\leq \sum_{k=1}^{N-1} \frac{1}{\eta_k^2} \\ &= \frac{1}{\eta_1^2} + \sum_{1\leq k\leq N-2, k\in\mathcal{B}} \frac{1}{\eta_{k+1}^2} + \sum_{1\leq k\leq N-2, k\notin\mathcal{B}} \frac{1}{\eta_{k+1}^2}.\end{aligned}$$

Recall that we have $\eta_{k+1} = \hat{\eta}_k$ if $k \in \mathcal{B}$ and $\eta_{k+1} = \hat{\eta}_k/\beta$ otherwise. Hence, we further have

$$\begin{aligned}\sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{1}{\hat{\eta}_k^2} &\leq \frac{1}{\eta_1^2} + \sum_{1\leq k\leq N-2, k\in\mathcal{B}} \frac{1}{\eta_{k+1}^2} + \sum_{1\leq k\leq N-2, k\notin\mathcal{B}} \frac{1}{\eta_{k+1}^2} \\ &= \frac{1}{\eta_1^2} + \sum_{1\leq k\leq N-2, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2} + \sum_{1\leq k\leq N-2, k\notin\mathcal{B}} \frac{\beta^2}{\hat{\eta}_k^2} \\ &\leq \frac{1}{\eta_1^2} + \sum_{1\leq k\leq N-1, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2} + \sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{\beta^2}{\hat{\eta}_k^2}.\end{aligned}$$

By moving the last term to the left-hand side and dividing both sides by $1-\beta^2$, we obtain

$$\sum_{1\leq k\leq N-1, k\notin\mathcal{B}} \frac{1}{\hat{\eta}_k^2} \leq \frac{1}{1-\beta^2}\left(\frac{1}{\eta_1^2} + \sum_{1\leq k\leq N-1, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2}\right). \tag{72}$$

Furthermore, since $\eta_1 \geq \hat{\eta}_0$, we have $\frac{1}{\eta_1^2} \leq \frac{1}{\hat{\eta}_0^2}$. Hence, by combining (71) and (72), we get

$$\sum_{k=0}^{N-1} \frac{1}{\hat{\eta}_k^2} \leq \frac{2-\beta^2}{1-\beta^2}\left(\frac{1}{\hat{\eta}_0^2} + \sum_{1\leq k\leq N-1, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2}\right) \leq \frac{2-\beta^2}{(1-\beta^2)\sigma_0^2} + \frac{2-\beta^2}{1-\beta^2}\sum_{0\leq k\leq N-1, k\in\mathcal{B}} \frac{1}{\hat{\eta}_k^2}, \tag{73}$$

where in the last inequality we used the fact that $\hat{\eta}_k = \sigma_0$ if $0 \notin \mathcal{B}$. Finally, (16) follows from Lemma 1 and (73).

C.2 The computational cost of Euclidean projection

Recall that $\mathcal{Z} \triangleq \{\mathbf{B} \in \mathbb{S}_+^d : 0 \preceq \mathbf{B} \preceq L_1\mathbf{I}\}$. As described in [34, Section D.1], the Euclidean projection on $\mathcal{Z}$ has a closed form solution. Specifically, Given the input $\mathbf{A} \in \mathbb{S}^d$, we first need to perform the eigendecomposition $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{V}$ is an orthogonal matrix and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix. Then the Euclidean projection of $\mathbf{A}$ onto $\mathcal{Z}$ is given by $\mathbf{V}\hat{\mathbf{\Lambda}}\mathbf{V}^\top$, where $\hat{\mathbf{\Lambda}}$ is a diagonal matrix with the diagonals being $\hat{\lambda}_k = \min\{L_1, \max\{0, \lambda_k\}\}$ for $1 \leq k \leq d$. Since the eigendecomposition requires $\mathcal{O}(d^3)$ arithmetic operations in general, the cost of computing the Euclidean projection can be prohibitive.

Algorithm 2 Projection-Free Online Learning

1: **Input:** Initial point $\mathbf{w}_0 \in \mathcal{B}_R(0)$, step size $\rho > 0$, $\delta > 0$
2: **for** $t = 0, 1, \ldots T-1$ **do**
3: Query the oracle $(\gamma_t, \mathbf{s}_t) \leftarrow \mathsf{SEP}(\mathbf{w}_t; \delta_t)$
4: **if** $\gamma_t \leq 1$ **then** *# Case I: we have* $\mathbf{w}_t \in \mathcal{C}$
5: Set $\mathbf{x}_t \leftarrow \mathbf{w}_t$ and play the action $\mathbf{x}_t$
6: Receive the loss $\ell_t(\mathbf{x}_t)$ and the gradient $\mathbf{g}_t = \nabla \ell_t(\mathbf{x}_t)$
7: Set $\tilde{\mathbf{g}}_t \leftarrow \mathbf{g}_t$
8: **else** *# Case II: we have* $\mathbf{w}_t/\gamma_t \in \mathcal{C}$
9: Set $\mathbf{x}_t \leftarrow \mathbf{w}_t/\gamma_t$ and play the action $\mathbf{x}_t$
10: Receive the loss $\ell_t(\mathbf{x}_t)$ and the gradient $\mathbf{g}_t = \nabla \ell_t(\mathbf{x}_t)$
11: Set $\tilde{\mathbf{g}}_t \leftarrow \mathbf{g}_t + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t \rangle\} \mathbf{s}_t$
12: **end if**
13: Update $\mathbf{w}_{t+1} \leftarrow \frac{R}{\max\{\|\mathbf{w}_t - \rho \tilde{\mathbf{g}}_t\|_2, R\}}(\mathbf{w}_t - \rho \tilde{\mathbf{g}}_t)$ *# Euclidean projection onto* $\mathcal{B}_R(0)$
14: **end for**

C.3 Online Learning with an Approximate Separation Oracle

To set the stage for our Hessian approximation matrix update, we first describe a projection-free online learning algorithm in a general setup. Specifically, the online learning protocol is as follows: For rounds $t = 0, 1, \ldots, T-1$, a learner chooses an action $\mathbf{x}_t \in \mathcal{C}$ from a convex set $\mathcal{C}$ and then observes a loss function $\ell_t : \mathbb{R}^n \to \mathbb{R}$. We measure the performance of an online learning algorithm by the dynamic regret [39, 42] defined by

$$\text{D-Reg}_T(\mathbf{u}_1, \ldots, \mathbf{u}_{T-1}) \triangleq \sum_{t=0}^{T-1} \ell_t(\mathbf{x}_t) - \sum_{t=0}^{T-1} \ell_t(\mathbf{u}_t),$$

where $\{\mathbf{u}_t\}_{t=1}^{T}$ is a sequence of comparators. Moreover, we assume that the convex set $\mathcal{C}$ is contained in the Euclidean ball $\mathcal{B}_R(0)$ for some $R > 0$, and we assume $0 \in \mathcal{C}$ without loss of generality.

Most existing online learning algorithms are projection-based, that is, they require computing the Euclidean projection on the action set $\mathcal{C}$. However, as we have seen in Section C.2, computing the projection is computationally costly in our setting. Inspired by the work in [40], we will describe an online learning algorithm that relies on an approximate separation oracle defined in Definition 3.

Definition 3. *The oracle* $\mathsf{SEP}(\mathbf{w}; \delta)$ *takes* $\mathbf{w} \in \mathcal{B}_R(0)$ *and* $\delta > 0$ *as input and returns a scalar* $\gamma > 0$ *and a vector* $\mathbf{s} \in \mathbb{R}^n$ *with one of the following possible outcomes:*

- *Case I:* $\gamma \leq 1$ *which implies that* $\mathbf{w} \in \mathcal{C}$*;*
- *Case II:* $\gamma > 1$ *which implies that* $\mathbf{w}/\gamma \in \mathcal{C}$ *and* $\langle \mathbf{s}, \mathbf{w} - \mathbf{x} \rangle \geq \gamma - 1 - \delta \quad \forall \mathbf{x} \in \mathcal{C}$.

In summary, the oracle $\mathsf{SEP}(\mathbf{w}; \delta)$ has two possible outcomes: it either certifies that $\mathbf{w}$ is feasible, i.e., $\mathbf{w} \in \mathcal{C}$, or it produces a scaled version of $\mathbf{w}$ that is in $\mathcal{C}$ and gives an approximate separating hyperplane between $\mathbf{w}$ and the set $\mathcal{C}$.

The full algorithm is shown in Algorithm 2. The key idea here is to introduce surrogate loss functions $\tilde{\ell}_t(\mathbf{w}) = \langle \tilde{\mathbf{g}}_t, \mathbf{w} \rangle$ on the larger set $\mathcal{B}_R(0)$ for $0 \leq t \leq T-1$, where $\tilde{\mathbf{g}}_t$ is the surrogate gradient to be defined later. On a high level, we will run online projected gradient descent with $\tilde{\ell}_t(\mathbf{w})$ to update the auxiliary iterates $\{\mathbf{w}_t\}_{t \geq 0}$ (note that the projection on $\mathcal{B}_R(0)$ is easy to compute), and then produce the actions $\{\mathbf{x}_t\}_{t \geq 0}$ for the original problem by calling the $\mathsf{SEP}(\mathbf{w}_t; \delta)$ oracle in Definition 3. The follow lemma shows that the immediate regret $\tilde{\ell}_t(\mathbf{w}_t) - \tilde{\ell}_t(\mathbf{x})$ can serve as an upper bound on $\ell_t(\mathbf{x}_t) - \ell_t(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{C}$.

Lemma 10. *Let* $\{\mathbf{x}_t\}_{t=0}^{T-1}$ *be the iterates generated by Algorithm 2. Then we have* $\mathbf{x}_t \in \mathcal{C}$ *for* $t = 0, 1, \ldots, T-1$*. Also, for any* $\mathbf{x} \in \mathcal{C}$*, we have*

$$\begin{aligned}
\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle &\leq \langle \tilde{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{x} \rangle + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t \rangle\} \delta_t && (74)\\
&\leq \frac{1}{2\rho}\|\mathbf{w}_t - \mathbf{x}\|_2^2 - \frac{1}{2\rho}\|\mathbf{w}_{t+1} - \mathbf{x}\|_2^2 + \frac{\rho}{2}\|\tilde{\mathbf{g}}_t\|_2^2 + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t \rangle\} \delta_t, && (75)
\end{aligned}$$

and

$$\|\tilde{\mathbf{g}}_t\| \leq \|\mathbf{g}_t\| + |\langle \mathbf{g}_t, \mathbf{x}_t \rangle| \|\mathbf{s}_t\|. \quad (76)$$

Subroutine 2 Online Learning Guided Hessian Approximation Update

Input: Initial matrix $\mathbf{B}_0 \in \mathbb{S}^d$ s.t. $0 \preceq \mathbf{B}_0 \preceq L_1\mathbf{I}$, step size $\rho > 0$, $\delta > 0$, $\{q_t\}_{t=1}^{T-1}$
Initialize: set $\mathbf{W}_0 \leftarrow \frac{2}{L_1}(\mathbf{B}_0 - \frac{L_1}{2}\mathbf{I})$, $\mathbf{G}_0 \leftarrow \frac{2}{L_1}\nabla \ell_0(\mathbf{B}_0)$ and $\tilde{\mathbf{G}}_0 \leftarrow \mathbf{G}_0$
for $t = 1, \dots, T-1$ **do**
  Query the oracle $(\gamma_t, \mathbf{S}_t) \leftarrow \mathsf{SEP}(\mathbf{W}_t; \delta_t, q_t)$
  if $\gamma_t \le 1$ **then** *# Case I*
    Set $\hat{\mathbf{B}}_t \leftarrow \mathbf{W}_t$ and $\mathbf{B}_t \leftarrow \frac{L_1}{2}\hat{\mathbf{B}}_t + \frac{L_1}{2}\mathbf{I}$
    Set $\mathbf{G}_t \leftarrow \frac{2}{L_1}\nabla \ell_t(\mathbf{B}_t)$ and $\tilde{\mathbf{G}}_t \leftarrow \mathbf{G}_t$
  else *# Case II*
    Set $\hat{\mathbf{B}}_t \leftarrow \mathbf{W}_t/\gamma_t$ and $\mathbf{B}_t \leftarrow \frac{L_1}{2}\hat{\mathbf{B}}_t + \frac{L_1}{2}\mathbf{I}$
    Set $\mathbf{G}_t \leftarrow \frac{2}{L_1}\nabla \ell_t(\mathbf{B}_t)$ and $\tilde{\mathbf{G}}_t \leftarrow \mathbf{G}_t + \max\{0, -\langle \mathbf{G}_t, \mathbf{B}_t\rangle\}\mathbf{S}_t$
  end if
  Update $\mathbf{W}_{t+1} \leftarrow \frac{\sqrt{d}}{\max\{\sqrt{d}, \|\mathbf{W}_t - \rho\tilde{\mathbf{G}}_t\|_F\}}(\mathbf{W}_t - \rho\tilde{\mathbf{G}}_t)$ *# Euclidean projection onto $\mathcal{B}_{\sqrt{d}}(0)$*
end for

Proof. By the definition of SEP in Definition 3, we can see that $\mathbf{x}_t \in \mathcal{C}$ for all $t = 1, \dots, T$. We now show that both (74) and (76) hold. We distinguish two cases depending on the outcomes of $\mathsf{SEP}(\mathbf{w}_t; \delta_t)$.

- If $\gamma_t \le 1$, then we have $\mathbf{x}_t = \mathbf{w}_t$ and $\tilde{\mathbf{g}}_t = \mathbf{g}_t$. In this case, (74) and (76) trivially hold.
- If $\gamma_t > 1$, then $\mathbf{x}_t = \mathbf{w}_t/\gamma_t$ and $\tilde{\mathbf{g}}_t = \mathbf{g}_t + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\mathbf{s}_t$. We can then write

$$
\begin{aligned}
\langle \tilde{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{x}\rangle &= \langle \mathbf{g}_t + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\mathbf{s}_t, \mathbf{w}_t - \mathbf{x}\rangle \\
&= \langle \mathbf{g}_t, \gamma_t\mathbf{x}_t - \mathbf{x}\rangle + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\langle \mathbf{s}_t, \mathbf{w}_t - \mathbf{x}\rangle \\
&\ge \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle + (\gamma_t - 1)\langle \mathbf{g}_t, \mathbf{x}_t\rangle + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}(\gamma_t - 1 - \delta_t) \\
&= \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle - \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\delta_t + (\gamma_t - 1)\max\{0, \langle \mathbf{g}_t, \mathbf{x}_t\rangle\} \\
&\ge \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle - \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\delta_t,
\end{aligned}
$$

  which leads to (74) after rearranging. Also, by the triangle inequality we obtain

$$
\|\tilde{\mathbf{g}}_t\| \le \|\mathbf{g}_t\| + \max\{0, -\langle \mathbf{g}_t, \mathbf{x}_t\rangle\}\|\mathbf{s}_t\| \le \|\mathbf{g}_t\| + |\langle \mathbf{g}_t, \mathbf{x}_t\rangle|\|\mathbf{s}_t\|,
$$

  which proves (76).

Finally, from the update rule of $\mathbf{w}_{t+1}$, for any $\mathbf{x} \in \mathcal{C} \subset \mathcal{B}_R(0)$ we have $\langle \mathbf{w}_t - \rho\tilde{\mathbf{g}}_t - \mathbf{w}_{t+1}, \mathbf{w}_{t+1} - \mathbf{x}\rangle \ge 0$, which further implies that

$$
\begin{aligned}
\langle \tilde{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{x}\rangle &\le \langle \tilde{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{w}_{t+1}\rangle + \frac{1}{\rho}\langle \mathbf{w}_t - \mathbf{w}_{t+1}, \mathbf{w}_{t+1} - \mathbf{x}\rangle && (77) \\
&= \langle \tilde{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{w}_{t+1}\rangle + \frac{1}{2\rho}\|\mathbf{w}_t - \mathbf{x}\|_2^2 - \frac{1}{2\rho}\|\mathbf{w}_{t+1} - \mathbf{x}\|_2^2 - \frac{1}{2\rho}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2 && (78) \\
&\le \frac{1}{2\rho}\|\mathbf{w}_t - \mathbf{x}\|_2^2 - \frac{1}{2\rho}\|\mathbf{w}_{t+1} - \mathbf{x}\|_2^2 + \frac{\rho}{2}\|\tilde{\mathbf{g}}_t\|_2^2. && (79)
\end{aligned}
$$

Combining (74) and (79) leads to (75). □

C.4 Projection-free Hessian Approximation Update

Now we are ready to describe our Hessian approximation matrix update, which is an specific instantiation of the general projection-free online learning algorithm described in Algorithm 2. The full algorithm is described in Subroutine 2.

Recall that $\mathcal{Z} = \{\mathbf{B} \in \mathbb{S}^d_+ : 0 \preceq \mathbf{B} \preceq L_1\mathbf{I}\}$ in our online learning problem in Section 3.2. Since the projection-free scheme in Subroutine 2 requires the set $\mathcal{C}$ to contain the origin, we consider the transform $\hat{\mathbf{B}} \triangleq \frac{2}{L_1}\left(\mathbf{B} - \frac{L_1}{2}\mathbf{I}\right)$ and define $\hat{\mathcal{Z}} \triangleq \{\hat{\mathbf{B}} \in \mathbb{S}^d : -\mathbf{I} \preceq \hat{\mathbf{B}} \preceq \mathbf{I}\} = \{\hat{\mathbf{B}} \in \mathbb{S}^d : \|\hat{\mathbf{B}}\|_{\mathrm{op}} \le 1\}$. We note that $0 \in \hat{\mathcal{Z}}$ and $\hat{\mathcal{Z}} \subset \mathcal{B}_{\sqrt{d}}(0) = \{\mathbf{W} \in \mathbb{S}^d : \|\mathbf{W}\|_F \le \sqrt{d}\}$. Moreover, we can see that the approximate separation oracle $\mathsf{SEP}(\mathbf{W}; \delta, q)$ defined in Definition 2 corresponds to the oracle in Definition 3. We defer the specific implementation details to Section E.2.

D Proof of Theorem 1

Regarding the choices of the hyper-parameters, we consider Algorithm 1 with the line search scheme in Subroutine 1, where $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 + \alpha_2 < 1$ and $\beta \in (0,1)$, and with the Hessian approximation update in Subroutine 2, where $\rho = \frac{1}{128}$, $q_t = p/2.5(t+1)\log^2(t+1)$ for $t \geq 1$, and $\delta_t = 1/(\sqrt{t+2}\ln(t+2))$ for $t \geq 0$. In the following, we first provide a proof sketch of Theorem 1. The complete proofs of the lemmas shown below will be provided in the subsequent sections.

Proof Sketch. To begin with, throughout the proof, we assume that every call of the SEP oracle in Definition 2 is successful during the execution of Algorithm 1. Indeed, by using the union bound, we can bound the failure probability by $\sum_{t=1}^{T-1} q_t \leq \frac{p}{2.5}\sum_{t=2}^{\infty} \frac{1}{t\log^2 t} \leq p$. In particular, we note that Subroutine 2 ensures that $0 \preceq \mathbf{B}_k \preceq L_1 \mathbf{I}$ for any $k \geq 0$.

We first prove Part (a) of Theorem 1, which relies on the following lemma.

Lemma 11. *For $k \in \mathcal{B}$, we have $\ell_k(\mathbf{B}_k) \triangleq \frac{\|\mathbf{w}_k - \mathbf{B}_k \mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2} \leq L_1^2$.*

We combine Lemma 2 and Lemma 11 to derive

$$\sum_{k=0}^{N-1} \frac{1}{\hat{\eta}_k^2} \leq \frac{2-\beta^2}{(1-\beta^2)\sigma_0^2} + \frac{2-\beta^2}{(1-\beta^2)\alpha_2^2\beta^2} \sum_{k\in\mathcal{B}} \frac{\|\mathbf{w}_k - \mathbf{B}_k\mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2} \leq \frac{2-\beta^2}{(1-\beta^2)\sigma_0^2} + \frac{(2-\beta^2)L_1^2}{(1-\beta^2)\alpha_2^2\beta^2} N.$$

By further using (15) and the elementary inequality that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we obtain

$$f(\mathbf{x}_N) - f(\mathbf{x}^*) \leq \frac{C_4 L_1 \|\mathbf{z}_0 - \mathbf{x}^*\|^2}{N^2} + \frac{C_5 \|\mathbf{z}_0 - \mathbf{x}^*\|^2}{\sigma_0 N^{2.5}}, \tag{80}$$

where $C_4 = C_1\sqrt{\frac{2-\beta^2}{(1-\beta^2)\sigma_0^2} + \frac{(2-\beta^2)}{(1-\beta^2)\alpha_2^2\beta^2}}$ and $C_5 = C_1\sqrt{\frac{2-\beta^2}{(1-\beta^2)\sigma_0^2}}$

Next, we divide the proof of Part (b) of Theorem 1 into the following steps.

Step 1: We first use regret analysis to control the cumulative loss $\sum_{t=0}^{T-1} \ell_t(\mathbf{B}_t)$ incurred by our online learning algorithm in Subroutine 2. In particular, we prove a dynamic regret bound, where we compare the cumulative loss of our algorithm against the one achieved by the sequence $\{\mathbf{H}_t\}_{t=0}^{T-1}$.

Lemma 12. *We have*

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{B}_t) \leq 256\|\mathbf{B}_0 - \mathbf{H}_0\|_F^2 + 4\sum_{t=0}^{T-1} \ell_t(\mathbf{H}_t) + 2L_1^2 \sum_{t=0}^{T-1} \delta_t^2 + 512 L_1 \sqrt{d} \sum_{t=0}^{T-1} \|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F,$$

where $\mathbf{H}_t \triangleq \nabla^2 f(\mathbf{y}_t)$.

Step 2: In light of Lemma 12, it suffices to upper bound the cumulative loss $\sum_{t=0}^{T-1} \ell_t(\mathbf{H}_t)$ and the path-length $\sum_{t=0}^{T-1} \|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F$ in the following lemma. To achieve this, we use the stability properties of our algorithm in (22) and Lemma 8, which is most technical part of the proof.

Lemma 13. *We have*

$$\sum_{t=0}^{T-1} \ell_t(\mathbf{H}_t) \leq \frac{C_3}{4} L_2^2 \|\mathbf{z}_0 - \mathbf{x}^*\|^2 \quad \text{and} \quad \sum_{t=0}^{T-1} \|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F \leq C_2 \sqrt{d} L_2 \left(1 + \log \frac{A_N}{A_1}\right) \|\mathbf{z}_0 - \mathbf{x}^*\|, \tag{81}$$

where C_2 is defined in (55) and $C_3 = \frac{(1+\alpha_1)^2}{\beta^2(1-\alpha_1)^2(1-\sigma^2)}$.

Step 3: Thus, we obtain an upper bound on $\sum_{t=0}^{T-1} \ell_t(\mathbf{B}_t)$ by combining Lemma 12 and Lemma 13. Finally, in the following lemma, we prove an upper bound on $\frac{1}{A_N}$ by further using Lemma 2 and Proposition 1.

Lemma 14. *We have*

$$\frac{1}{A_N} \leq \frac{1}{N^{2.5}}\left(M + C_{10}L_1L_2d\|\mathbf{z}_0 - \mathbf{x}^*\|\log^+\left(\frac{\max\{\frac{L_1}{\alpha_2\beta}, \frac{1}{\sigma_0}\}N^{2.5}}{\sqrt{M}}\right)\right)^{\frac{1}{2}},$$

where we define $\log^+(x) \triangleq \max\{\log(x), 0\}$,

$$M = \frac{C_6}{\sigma_0^2} + C_7L_1^2 + C_8\|\mathbf{B}_0 - \mathbf{H}_0\|_F^2 + C_9L_2^2\|\mathbf{z}_0 - \mathbf{x}^*\|^2 + C_{10}L_1L_2d\|\mathbf{z}_0 - \mathbf{x}^*\|,$$

and C_i $(i = 6, \ldots, 10)$ *are absolute constants given by*

$$C_6 = \frac{4C_1^2(2-\beta^2)}{1-\beta^2},\ C_7 = \frac{5C_6}{\alpha_2^2\beta^2},\ C_8 = \frac{256C_6}{\alpha_2^2\beta^2},\ C_9 = \frac{C_3C_6}{\alpha_2^2\beta^2},\ C_{10} = \frac{512C_2C_6}{\alpha_2^2\beta^2}.$$

Therefore, Part (b) of Theorem 1 immediately follows from Proposition 1. □

In the remaining of this section, we present the proofs for the above lemmas that we used to prove the results in Theorem 1.

D.1 Proof of Lemma 11

Recall that $\mathbf{w}_k \triangleq \nabla f(\tilde{\mathbf{x}}_{k+1}) - \nabla f(\mathbf{y}_k)$ and $\mathbf{s}_k \triangleq \tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k$ for $k \in \mathcal{B}$. We can write $\nabla f(\tilde{\mathbf{x}}_{k+1}) - \nabla f(\mathbf{y}_k) = \bar{\mathbf{H}}_k(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k)$ by using the fundamental theorem of calculus, where $\bar{\mathbf{H}}_k = \int_0^1 \nabla^2 f(t\tilde{\mathbf{x}}_{k+1} + (1-t)\mathbf{y}_k)\,dt$. Since we have $0 \preceq \nabla^2 f(\mathbf{x}) \preceq L_1\mathbf{I}$ for all $\mathbf{x} \in \mathbb{R}^d$ by Assumption 1, it implies that $0 \preceq \bar{\mathbf{H}}_k \preceq L_1\mathbf{I}$. Moreover, since $0 \preceq \mathbf{B}_k \preceq L_1\mathbf{I}$, we further have $-L_1\mathbf{I} \preceq \bar{\mathbf{H}}_k - \mathbf{B}_k \preceq L_1\mathbf{I}$, which yields $\|\bar{\mathbf{H}}_k - \mathbf{B}_k\|_{\mathrm{op}} \leq L_1$. Thus, we have

$$\|\mathbf{w}_k - \mathbf{B}_k\mathbf{s}_k\| = \|(\bar{\mathbf{H}}_k - \mathbf{B}_k)(\tilde{\mathbf{x}}_{k+1} - \mathbf{y}_k)\| \leq L_1\|\tilde{\mathbf{x}}_k - \mathbf{x}_k\|,$$

which proves that $\ell_k(\mathbf{B}_k) \leq L_1^2$.

D.2 Proof of Lemma 12

To prove Lemma 12, we first present the following lemma showing a smooth property of the loss function ℓ_k. The proof is similar to [34, Lemma 15].

Lemma 15. *For* $k \in \mathcal{B}$, *we have*

$$\nabla\ell_k(\mathbf{B}) = \frac{1}{\|\mathbf{s}_k\|^2}\left(-\mathbf{s}_k(\mathbf{w}_k - \mathbf{B}\mathbf{s}_k)^\mathsf{T} - (\mathbf{w}_k - \mathbf{B}\mathbf{s}_k)\mathbf{s}_k^\mathsf{T}\right). \tag{82}$$

Moreover, for any $\mathbf{B} \in \mathbb{S}^d$, *it holds that*

$$\|\nabla\ell_k(\mathbf{B})\|_F \leq \|\nabla\ell_k(\mathbf{B})\|_* \leq 2\sqrt{\ell_k(\mathbf{B})}, \tag{83}$$

where $\|\cdot\|_F$ *and* $\|\cdot\|_*$ *denote the Frobenius norm and the nuclear norm, respectively.*

Proof. It is straightforward to verify the expression in (82). The first inequality in (83) follows from the fact that $\|\mathbf{A}\|_F \leq \|\mathbf{A}\|_*$ for any matrix $\mathbf{A} \in \mathbb{S}^d$. For the second inequality, note that

$$\begin{aligned}\|\nabla\ell_k(\mathbf{B})\|_* &\leq \frac{1}{\|\mathbf{s}_k\|^2}\left(\|\mathbf{s}_k(\mathbf{w}_k - \mathbf{B}\mathbf{s}_k)^\mathsf{T}\|_* + \|(\mathbf{w}_k - \mathbf{B}\mathbf{s}_k)\mathbf{s}_k^\mathsf{T}\|_*\right)\\ &\leq \frac{2}{\|\mathbf{s}_k\|^2}\|\mathbf{w}_k - \mathbf{B}\mathbf{s}_k\|\|\mathbf{s}_k\| = \frac{2\|\mathbf{w}_k - \mathbf{B}\mathbf{s}_k\|}{\|\mathbf{s}_k\|} = 2\sqrt{\ell_k(\mathbf{B})},\end{aligned}$$

where in the first inequality we used the triangle inequality, and in the second inequality we used the fact that the rank-one matrix $\mathbf{u}\mathbf{v}^\top$ has only one nonzero singular value $\|\mathbf{u}\|\|\mathbf{v}\|$. □

We will also need the following helper lemma.

Lemma 16. *If the real number* x *satisfies* $x \leq A + B\sqrt{x}$, *then we have* $x \leq 2A + B^2$.

Proof. From the assumption, we have

$$\left(\sqrt{x} - \frac{B}{2}\right)^2 \le A + \frac{B^2}{4}.$$

Hence, we obtain

$$x \le \left(\sqrt{A + \frac{B^2}{4}} + \frac{B}{2}\right)^2 \le 2A + B^2.$$

$\square$

Before proving Lemma 12, we also present the following lemma that bounds the loss in each round.

Lemma 17. *For any $\mathbf{H} \in \mathcal{Z}$, we have*

$$\ell_t(\mathbf{B}_t) \le 4\ell_t(\mathbf{H}) + 64L_1^2\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - 64L_1^2\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + 2L_1^2\delta_t^2.$$

Proof. By letting $\mathbf{x}_t = \hat{\mathbf{B}}_t$, $\mathbf{x} = \hat{\mathbf{H}} \triangleq \frac{2}{L_1}(\mathbf{H} - \frac{L_1}{2}\mathbf{I})$, $\mathbf{g}_t = \mathbf{G}_t \triangleq \frac{2}{L_1}\nabla\ell_t(\mathbf{B}_t)$, $\tilde{\mathbf{g}}_t = \tilde{\mathbf{G}}_t$, $\mathbf{w}_t = \mathbf{W}_t$ in Lemma 10, we obtain:

(i) $\hat{\mathbf{B}}_t \in \hat{\mathcal{Z}}$, which means that $\|\hat{\mathbf{B}}_t\|_{\mathrm{op}} \le 1$.

(ii) It holds that

$$\langle \mathbf{G}_t, \hat{\mathbf{B}}_t - \hat{\mathbf{H}}\rangle \le \frac{1}{2\rho}\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - \frac{1}{2\rho}\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + \frac{\rho}{2}\|\tilde{\mathbf{G}}_t\|_F^2 + \max\{0, -\langle \mathbf{G}_t, \hat{\mathbf{B}}_t\rangle\}\delta_t, \tag{84}$$

$$\|\tilde{\mathbf{G}}_t\|_F \le \|\mathbf{G}_t\|_F + |\langle \mathbf{G}_t, \hat{\mathbf{B}}_t\rangle|\|\mathbf{S}_t\|_F. \tag{85}$$

First, note that $\|\mathbf{S}_t\|_F \le 3$ by Definition 2 and $|\langle \mathbf{G}_t, \hat{\mathbf{B}}_t\rangle| \le \|\mathbf{G}_t\|_*\|\hat{\mathbf{B}}_t\|_{\mathrm{op}} \le \|\mathbf{G}_t\|_*$. Together with (85), we get

$$\|\tilde{\mathbf{G}}_t\|_F \le \|\mathbf{G}_t\|_F + 3\|\mathbf{G}_t\|_* \le 4\|\mathbf{G}_t\|_* \le \frac{16}{L_1}\sqrt{\ell_t(\mathbf{B}_t)}, \tag{86}$$

where we used the fact that $\mathbf{G}_t = \frac{2}{L_1}\nabla\ell_t(\mathbf{B}_t)$ and Lemma 15 in the last inequality. Furthermore, since ℓ_t is convex, we have

$$\ell_t(\mathbf{B}_t) - \ell_t(\mathbf{H}) \le \langle \nabla\ell_t(\mathbf{B}_t), \mathbf{B}_t - \mathbf{H}\rangle = \left(\frac{L_1}{2}\right)^2 \langle \mathbf{G}_t, \hat{\mathbf{B}}_t - \hat{\mathbf{H}}\rangle,$$

where we used $\mathbf{G}_t = \frac{2}{L_1}\nabla\ell_t(\mathbf{B}_t)$, $\hat{\mathbf{B}}_t \triangleq \frac{2}{L_1}(\mathbf{B}_t - \frac{L_1}{2}\mathbf{I})$, and $\hat{\mathbf{H}} \triangleq \frac{2}{L_1}(\mathbf{H} - \frac{L_1}{2}\mathbf{I})$. Therefore, by combining (84) and (86) we get

$$\begin{aligned}\ell_t(\mathbf{B}_t) - \ell_t(\mathbf{H}) &\le \frac{L_1^2}{8\rho}\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - \frac{L_1^2}{8\rho}\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + \frac{\rho}{8}L_1^2\|\tilde{\mathbf{G}}_t\|_F^2 + \frac{L_1^2}{4}\|\mathbf{G}_t\|_*\delta_t && (87)\\ &\le \frac{L_1^2}{8\rho}\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - \frac{L_1^2}{8\rho}\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + 32\rho\ell_t(\mathbf{B}_t) + L_1\sqrt{\ell_t(\mathbf{B}_t)}\delta_t. && (88)\end{aligned}$$

Note that $\ell_t(\mathbf{B}_t)$ appears on both sides of (88). By further applying Lemma 16, we obtain

$$\ell_t(\mathbf{B}_t) \le 2\ell_t(\mathbf{H}) + \frac{L_1^2}{4\rho}\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - \frac{L_1^2}{4\rho}\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + 64\rho\ell_t(\mathbf{B}_t) + L_1^2\delta_t^2.$$

Since $\rho = 1/128$, by rearranging and simplifying terms in the above inequality, we obtain

$$\ell_t(\mathbf{B}_t) \le 4\ell_t(\mathbf{H}) + 64L_1^2\|\mathbf{W}_t - \hat{\mathbf{H}}\|_F^2 - 64L_1^2\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}\|_F^2 + 2L_1^2\delta_t^2.$$

$\square$

Proof of Lemma 12. We let $\mathbf{H}_t = \nabla^2 f(\mathbf{y}_t)$ for $t = 0, 1, \ldots, T-1$. Thus, we get

$$\begin{aligned}\ell_t(\mathbf{B}_t) &\leq 4\ell_t(\mathbf{H}_t) + 64L_1^2\|\mathbf{W}_t - \hat{\mathbf{H}}_t\|_F^2 - 64L_1^2\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_t\|_F^2 + 2L_1^2\delta_t^2 \\ &= 4\ell_t(\mathbf{H}_t) + 64L_1^2\|\mathbf{W}_t - \hat{\mathbf{H}}_t\|_F^2 - 64L_1^2\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F^2 + 2L_1^2\delta_t^2 \\ &\quad + 64L_1^2\big(\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F^2 - \|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_t\|_F^2\big).\end{aligned}$$

Furthermore, note that

$$\begin{aligned}&\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F^2 - \|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_t\|_F^2 \\ &= (\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F + \|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_t\|_F)(\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F - \|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_t\|_F) \\ &\leq 4\sqrt{d}\|\hat{\mathbf{H}}_{t+1} - \hat{\mathbf{H}}_t\|_F = \frac{8\sqrt{d}}{L_1}\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F,\end{aligned}$$

where in the last inequality we used the fact that $\hat{\mathbf{H}}_t, \hat{\mathbf{H}}_{t+1}, \mathbf{W}_{t+1} \in \mathcal{B}_{\sqrt{d}}(0)$ and the triangle inequality. Therefore, we get

$$\ell_t(\mathbf{B}_t) \leq 4\ell_t(\mathbf{H}_t) + 64L_1^2\|\mathbf{W}_t - \hat{\mathbf{H}}_t\|_F^2 - 64L_1^2\|\mathbf{W}_{t+1} - \hat{\mathbf{H}}_{t+1}\|_F^2 + 2L_1^2\delta_t^2 + 512L_1\sqrt{d}\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F.$$

By summing the above inequality from $t = 0$ to $T-1$, we get

$$\sum_{t=0}^{T-1}\ell_t(\mathbf{B}_t) \leq 64L_1^2\|\mathbf{W}_0 - \hat{\mathbf{H}}_0\|_F^2 + 4\sum_{t=0}^{T-1}\ell_t(\mathbf{H}_t) + 2L_1^2\sum_{t=0}^{T-1}\delta_t^2 + 512L_1\sqrt{d}\sum_{t=0}^{T-1}\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F.$$

Finally, we use the fact that $\mathbf{W}_0 \triangleq \frac{2}{L_1}(\mathbf{B}_0 - \frac{L_1}{2}\mathbf{I})$, and $\hat{\mathbf{H}}_0 \triangleq \frac{2}{L_1}(\mathbf{H}_0 - \frac{L_1}{2}\mathbf{I})$ to obtain Lemma 12. □

D.3 Proof of Lemma 13

By Assumption 2, we have $\|\mathbf{w}_t - \mathbf{H}_t\mathbf{s}_t\| = \|\nabla f(\tilde{\mathbf{x}}_{t+1}) - \nabla f(\mathbf{y}_t) - \nabla f(\mathbf{y}_t)(\tilde{\mathbf{x}}_{t+1} - \mathbf{y}_t)\| \leq \frac{L_2}{2}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{y}_t\|^2$. Thus,

$$\ell_t(\mathbf{H}_t) = \frac{\|\mathbf{w}_t - \mathbf{H}_t\mathbf{s}_t\|^2}{\|\mathbf{s}_t\|^2} \leq \frac{L_2^2}{4}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{y}_t\|^2 \leq \frac{(1+\alpha_1)^2L_2^2}{4\beta^2(1-\alpha_1)^2}\|\hat{\mathbf{x}}_{t+1} - \mathbf{y}_t\|^2,$$

where we used Lemma 1 in the last inequality. Also, Since $a_k \geq \eta_k$ for all $k \geq 0$, by (22) we get

$$\sum_{k=0}^{N-1}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 \leq \sum_{k=0}^{N-1}\frac{a_k^2}{\eta_k^2}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 \leq \frac{1}{1-\sigma^2}\|\mathbf{z}_0 - \mathbf{x}^*\|^2.$$

Hence, we have

$$\begin{aligned}\sum_{t=0}^{T-1}\ell_t(\mathbf{H}_t) \leq \frac{(1+\alpha_1)^2L_2^2}{4\beta^2(1-\alpha_1)^2}\sum_{k\in\mathcal{B}}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 &\leq \frac{(1+\alpha_1)^2L_2^2}{4\beta^2(1-\alpha_1)^2}\sum_{k=0}^{N-1}\|\hat{\mathbf{x}}_{k+1} - \mathbf{y}_k\|^2 \\ &\leq \frac{(1+\alpha_1)^2L_2^2\|\mathbf{z}_0 - \mathbf{x}^*\|^2}{4\beta^2(1-\alpha_1)^2(1-\sigma^2)},\end{aligned}$$

which proves the first inequality in (81).

Furthermore, by Assumption 2, we have

$$\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F = \|\nabla^2 f(\mathbf{y}_{t+1}) - \nabla^2 f(\mathbf{y}_t)\|_F \leq \sqrt{d}\|\nabla^2 f(\mathbf{y}_{t+1}) - \nabla^2 f(\mathbf{y}_t)\|_{\mathrm{op}} \leq \sqrt{d}L_2\|\mathbf{y}_{t+1} - \mathbf{y}_t\|.$$

Hence, by using the triangle inequality, we can bound

$$\sum_{t=0}^{T-1}\|\mathbf{H}_{t+1} - \mathbf{H}_t\|_F \leq \sqrt{d}L_2\sum_{k=0}^{N-1}\|\mathbf{y}_{k+1} - \mathbf{y}_k\| \leq \sqrt{d}L_2C_2\left(1 + \log\frac{A_N}{A_1}\right)\|\mathbf{z}_0 - \mathbf{x}^*\|,$$

where we used Lemma 8 in the last inequality.

D.4 Proof of Lemma 14

We combine Lemma 12 and Lemma 13 to get

$$\sum_{k\in\mathcal{B}} \frac{\|\mathbf{w}_k - \mathbf{B}_k\mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2} = \sum_{t=0}^{T-1} \ell_t(\mathbf{B}_t) \le 256\|\mathbf{B}_0 - \mathbf{H}_0\|_F^2 + C_3 L_2^2\|\mathbf{z}_0 - \mathbf{x}^*\|^2 + 2L_1^2 \sum_{t=0}^{T-1}\delta_t^2 + 512 C_2 L_1 L_2 d\left(1 + \log\frac{A_N}{A_1}\right)\|\mathbf{z}_0 - \mathbf{x}^*\|.$$

Since $\delta_t = 1/(\sqrt{t+2}\ln(t+2))$, we have

$$\sum_{t=0}^{T-1}\delta_t^2 = \sum_{t=2}^{T+1}\frac{1}{t\ln^2 t} \le \frac{1}{2\ln^2 2} + \int_2^{T+1}\frac{1}{t\ln^2 t}dt = \frac{1}{2\ln^2 2} + \frac{1}{\ln 2} - \frac{1}{\ln(T+1)} \le 2.5.$$

Hence, it further follows from (15) and Lemma 2 that

$$\begin{aligned}\frac{N^5}{A_N^2} &\le 4C_1^2\sum_{k=0}^{N-1}\frac{1}{\hat{\eta}_k^2} \\ &\le \frac{4C_1^2(2-\beta^2)}{(1-\beta^2)\sigma_0^2} + \frac{4C_1^2(2-\beta^2)}{(1-\beta^2)\alpha_2^2\beta^2}\sum_{k\in\mathcal{B}}\frac{\|\mathbf{w}_k - \mathbf{B}_k\mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2} \\ &\le \frac{C_6}{\sigma_0^2} + C_7 L_1^2 + C_8\|\mathbf{B}_0 - \mathbf{H}_0\|_F^2 + C_9 L_2^2\|\mathbf{z}_0 - \mathbf{x}^*\|^2 \\ &\quad + C_{10}L_1L_2 d\left(1 + \log\frac{A_N}{A_1}\right)\|\mathbf{z}_0 - \mathbf{x}^*\|.\end{aligned}$$

To simplify the notation, define

$$M = \frac{C_6}{\sigma_0^2} + C_7 L_1^2 + C_8\|\mathbf{B}_0 - \mathbf{H}_0\|_F^2 + C_9 L_2^2\|\mathbf{z}_0 - \mathbf{x}^*\|^2 + C_{10}L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|.$$

Let A_N^* be the number that achieves the equality

$$\frac{N^5}{(A_N^*)^2} = M + C_{10}L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|\log\frac{A_N^*}{A_1}.$$

We can see that $A_N \ge A_N^*$. Thus, we instead try to construct a lower bound on A_N^*. If $A_N^* \le A_1$, then $\log(A_N^*/A_1) \le 0$ and furthermore

$$\frac{N^5}{(A_N^*)^2} \le M \quad \Rightarrow \quad A_N^* \ge \frac{1}{\sqrt{M}}N^{2.5}.$$

Otherwise, assume that $A_N^* > A_1$. Then $\log(A_N^*/A_1) > 0$ and we first show an upper bound on A_N^*:

$$\frac{N^5}{(A_N^*)^2} = M + C_8 L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|\log\frac{A_N^*}{A_1} \ge M \quad \Rightarrow \quad A_N^* \le \frac{1}{\sqrt{M}}N^{2.5}.$$

This in turn leads to a lower bound on A_N^*:

$$\frac{N^5}{(A_N^*)^2} = M + C_8 L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|\log\frac{A_N^*}{A_1} \le M + C_8 L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|\log\left(\frac{\max\{\frac{L_1}{\alpha_2\beta}, \frac{1}{\sigma_0}\}N^{2.5}}{\sqrt{M}}\right),$$

where we also used the fact that $A_1 = \hat{\eta}_1 \ge \min\{\sigma_0, \frac{\alpha_2\beta}{L_1}\}$. Thus, we get

$$\frac{1}{A_N} \le \frac{1}{A_N^*} \le \frac{1}{N^{2.5}}\left(M + C_{10}L_1L_2 d\|\mathbf{z}_0 - \mathbf{x}^*\|\log\left(\frac{\max\{\frac{L_1}{\alpha_2\beta}, \frac{1}{\sigma_0}\}N^{2.5}}{\sqrt{M}}\right)\right)^{\frac{1}{2}}.$$

Subroutine 3 $\mathsf{LinearSolver}(\mathbf{A}, \mathbf{b}; \alpha)$

Input: $\mathbf{A} \in \mathbb{S}^d_+$, $\mathbf{b} \in \mathbb{R}^d$, $0 < \alpha < 1$
Initialize: $\mathbf{s}_0 \leftarrow 0$, $\mathbf{r}_0 \leftarrow \mathbf{b} - \mathbf{A}\mathbf{s}_0$, $\mathbf{p}_0 \leftarrow \mathbf{r}_0$
for $k = 0, 1, \ldots$ **do**
  if $\|\mathbf{r}_k\|_2 \leq \alpha \|\mathbf{s}_k\|_2$ **then**
    Return $\mathbf{s}_k$
  end if
  $\alpha_k \leftarrow \langle \mathbf{r}_k, \mathbf{A}\mathbf{r}_k \rangle / \langle \mathbf{A}\mathbf{p}_k, \mathbf{A}\mathbf{p}_k \rangle$
  $\mathbf{s}_{k+1} \leftarrow \mathbf{s}_k + \alpha_k \mathbf{p}_k$
  $\mathbf{r}_{k+1} \leftarrow \mathbf{r}_k - \alpha_k \mathbf{A}\mathbf{p}_k$
  Compute and store $\mathbf{A}\mathbf{r}_{k+1}$
  $\beta_k \leftarrow \langle \mathbf{r}_{k+1}, \mathbf{A}\mathbf{r}_{k+1} \rangle / \langle \mathbf{r}_k, \mathbf{A}\mathbf{r}_k \rangle$
  $\mathbf{p}_{k+1} \leftarrow \mathbf{r}_{k+1} + \beta_k \mathbf{p}_k$
  Compute and store $\mathbf{A}\mathbf{p}_{k+1} \leftarrow \mathbf{A}\mathbf{r}_{k+1} + \beta_k \mathbf{A}\mathbf{p}_k$
end for

E Characterizing the Computational Cost

In this section, we first specify the implementation details of the LinearSolver oracle in Definition 1 and the SEP oracle in Definition 2. Then in Section E.3, we present the proof of Theorem 2.

E.1 Implementation of the LinearSolver Oracle

We implement the LinearSolver oracle by running the conjugate residual (CR) method [38] to solve the linear system $\mathbf{A}\mathbf{s} = \mathbf{b}$. In particular, we initialize the CT method with $\mathbf{s}_0 = 0$ and returns the iterate $\mathbf{s}_k$ once we have $\|\mathbf{A}\mathbf{s}_k - \mathbf{b}\| \leq \alpha \|\mathbf{s}_k\|$.

The following lemma provides the convergence guarantee of the CR method, which will be later used in the proof of Theorem 2.

Lemma 18. *Let $\mathbf{s}^*$ be any optimal solution of $\mathbf{A}\mathbf{s}^* = \mathbf{b}$ and let $\{\mathbf{s}_k\}$ be the iterates generated by Subourtine 3. Then we have*

$$\|\mathbf{r}_k\|_2 = \|\mathbf{A}\mathbf{s}_k - \mathbf{b}\|_2 \leq \frac{\lambda_{\max}(\mathbf{A})\|\mathbf{s}^*\|_2}{(k+1)^2}.$$

E.2 Implementation of SEP Oracle

We implement the SEP oracle in Definition 2 by running he classical Lanczos method, with a random start, where the initial vector is chosen randomly and uniformly from the unit sphere (see, e.g., [45, 46]). For completeness, the full algorithm is shown in Subroutine 4.

To prove the correctness of our algorithm, we first recall a classical result in [41] on the convergence behavior of the Lanczos method.

Proposition 3 ([41, Theorem 4.2])**.** *Consider a symmetric matrix $\mathbf{W}$ and let $\lambda_1(\mathbf{W})$ and $\lambda_d(\mathbf{W})$ denote its largest and smallest eigenvalues, respectively. Then after k iterations of the Lanczos method with a random start, we find unit vectors $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(d)}$ such that*

$$\mathbb{P}(\langle \mathbf{W}\mathbf{u}^{(1)}, \mathbf{u}^{(1)} \rangle \leq \lambda_1(\mathbf{W}) - \epsilon(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W}))) \leq 1.648\sqrt{d}e^{-\sqrt{\epsilon}(2k-1)},$$
$$\mathbb{P}(\langle \mathbf{W}\mathbf{u}^{(d)}, \mathbf{u}^{(d)} \rangle \geq \lambda_d(\mathbf{W}) + \epsilon(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W}))) \leq 1.648\sqrt{d}e^{-\sqrt{\epsilon}(2k-1)},$$

As a corollary, to ensure that, with probability at least $1 - q$,

$$\langle \mathbf{W}\mathbf{u}^{(1)}, \mathbf{u}^{(1)} \rangle > \lambda_1(\mathbf{W}) - \epsilon(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})) \text{ and } \langle \mathbf{W}\mathbf{u}^{(d)}, \mathbf{u}^{(d)} \rangle < \lambda_n(\mathbf{W}) + \epsilon(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})),$$

the number of iterations can be bounded by $\lceil \frac{1}{4}\epsilon^{-1/2}\log(11d/q^2) + \frac{1}{2} \rceil$.

Lemma 19. *Let γ and $\mathbf{S}$ be the output of $\mathsf{SEP}(\mathbf{W}; \delta, q)$ in Subroutine 4. Then with probability at least $1 - q$, they satisfy one of the following properties:*

- *Case I: $\gamma \leq 1$, then we have $\|\mathbf{W}\|_{\mathrm{op}} \leq 1$;*

Subroutine 4 SEP$(\mathbf{W}; \delta, q)$

Input: $\mathbf{W} \in \mathbb{S}^d$, $\delta > 0$, $q \in (0,1)$
Initialize: sample $\mathbf{v}_1 \in \mathbb{R}^d$ uniformly from the unit sphere, $\beta_1 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$
Set the number of iterations $N_1 \leftarrow \min\left\{\left\lceil \log \frac{11d}{q^2} + \frac{1}{2} \right\rceil, d\right\}$
for $k = 1, \ldots, N_1$ **do**
  Set $\mathbf{w}_k \leftarrow \mathbf{W}\mathbf{v}_k - \beta_k \mathbf{v}_{k-1}$
  Set $\alpha_k \leftarrow \langle \mathbf{w}_k, \mathbf{v}_k \rangle$ and $\mathbf{w}_k \leftarrow \mathbf{w}_k - \alpha_k \mathbf{v}_k$
  Set $\beta_{k+1} \leftarrow \|\mathbf{w}_k\|$ and $\mathbf{v}_{k+1} \leftarrow \mathbf{w}_k / \beta_{k+1}$
end for
Form a tridiagonal matrix $\mathbf{T} \leftarrow \mathsf{tridiag}(\beta_{2:N_1}, \alpha_{1:N_1}, \beta_{2:N_1})$
 # Use the tridiagonal structure to compute eigenvectors of $\mathbf{T}$
Compute $(\hat{\lambda}_1, \mathbf{z}^{(1)}) \leftarrow \mathsf{MaxEvec}(\mathbf{T})$ and $(\hat{\lambda}_d, \mathbf{z}^{(d)}) \leftarrow \mathsf{MinEvec}(\mathbf{T})$
Set $\mathbf{u}^{(1)} \leftarrow \sum_{k=1}^{N_1} z_k^{(1)} \mathbf{v}_k$ and $\mathbf{u}^{(d)} \leftarrow \sum_{k=1}^{N_1} z_k^{(d)} \mathbf{v}_k$
Set $\hat{\lambda}_{\max} \leftarrow \max\{\hat{\lambda}_1, -\hat{\lambda}_d\}$
if $\hat{\lambda}_{\max} \leq 1/2$ **then** *# Case I: $\gamma \leq 1$, which implies $\|\mathbf{W}\|_{\mathrm{op}} \leq 1$*
  Return $\gamma = 2\hat{\lambda}_{\max}$ and $\mathbf{S} = 0$
else if $\hat{\lambda}_{\max} \geq 2$ **then** *# Case II: $\gamma > 1$ and $\mathbf{S}$ defines a separating hyperplane*
  if $\hat{\lambda}_1 > -\hat{\lambda}_d$ **then**
    Return $\gamma = 2\hat{\lambda}_{\max}$ and $\mathbf{S} = 3\mathbf{u}^{(1)}(\mathbf{u}^{(1)})^\top$
  else
    Return $\gamma = 2\hat{\lambda}_{\max}$ and $\mathbf{S} = -3\mathbf{u}^{(d)}(\mathbf{u}^{(d)})^\top$
  end if
else *# $\frac{1}{2} < \hat{\lambda}_{\max} < 2$*
  Set the number of iterations $N_2 \leftarrow \min\left\{\left\lceil \frac{1}{4\sqrt{2\delta}} \log \frac{44d}{q^2} + \frac{1}{2} \right\rceil, d\right\}$
  for $k = N_1 + 1, \ldots, N_2$ **do**
    Set $\mathbf{w}_k \leftarrow \mathbf{W}\mathbf{v}_k - \beta_k \mathbf{v}_{k-1}$
    Set $\alpha_k \leftarrow \langle \mathbf{w}_k, \mathbf{v}_k \rangle$ and $\mathbf{w}_k \leftarrow \mathbf{w}_k - \alpha_k \mathbf{v}_k$
    Set $\beta_{k+1} \leftarrow \|\mathbf{w}_k\|$ and $\mathbf{v}_{k+1} \leftarrow \mathbf{w}_k / \beta_{k+1}$
  end for
  Form a tridiagonal matrix $\mathbf{T} \leftarrow \mathsf{tridiag}(\beta_{2:N_2}, \alpha_{1:N_2}, \beta_{2:N_2})$
  Compute $(\tilde{\lambda}_1, \tilde{\mathbf{z}}^{(1)}) \leftarrow \mathsf{MaxEvec}(\mathbf{T})$ and $(\tilde{\lambda}_d, \tilde{\mathbf{z}}^{(d)}) \leftarrow \mathsf{MinEvec}(\mathbf{T})$
  Set $\tilde{\mathbf{u}}^{(1)} \leftarrow \sum_{k=1}^{N_2} \tilde{z}_k^{(1)} \mathbf{v}_k$ and $\tilde{\mathbf{u}}^{(d)} \leftarrow \sum_{k=1}^{N_2} \tilde{z}_k^{(d)} \mathbf{v}_k$
  Lanczos method
  Set $\tilde{\lambda}_{\max} = \max\{\tilde{\lambda}_1, -\tilde{\lambda}_d\}$
  if $\tilde{\lambda}_{\max} \leq 1 - \delta$ **then**
    Return $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = 0$
  else if $\tilde{\lambda}_1 \geq -\tilde{\lambda}_d$ **then**
    Return $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = \tilde{\mathbf{u}}^{(1)}(\tilde{\mathbf{u}}^{(1)})^\top$
  else
    Return $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = -\tilde{\mathbf{u}}^{(d)}(\tilde{\mathbf{u}}^{(d)})^\top$
  end if
end if

- *Case II: $\gamma > 1$, then we have $\|\mathbf{W}/\gamma\|_{\mathrm{op}} \leq 1$, $\|\mathbf{S}\|_F = 3$ and $\langle \mathbf{S}, \mathbf{W} - \hat{\mathbf{B}} \rangle \geq \gamma - 1$ for any $\hat{\mathbf{B}}$ such that $\|\hat{\mathbf{B}}\|_{\mathrm{op}} \leq 1$.*

Proof. Note that in Subroutine 4, we first run the Lanczos method for $\left\lceil \epsilon^{-1/2} \log \frac{11d}{q^2} + \frac{1}{2} \right\rceil$ iterations, where $\epsilon = \frac{1}{4}$. Thus, by Proposition 3, with probability at least $1 - q/2$ we have

$$\hat{\lambda}_1 \triangleq \langle \mathbf{W}\mathbf{u}^{(1)}, \mathbf{u}^{(1)} \rangle \geq \lambda_1(\mathbf{W}) - \frac{1}{4}(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})), \tag{89}$$

$$\hat{\lambda}_d \triangleq \langle \mathbf{W}\mathbf{u}^{(d)}, \mathbf{u}^{(d)} \rangle \leq \lambda_d(\mathbf{W}) + \frac{1}{4}(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})). \tag{90}$$

Combining (89) and (90), we get

$$\frac{1}{2}(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})) \leq \hat{\lambda}_1 - \hat{\lambda}_d \quad \Rightarrow \quad \lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W}) \leq 2(\hat{\lambda}_1 - \hat{\lambda}_d).$$

By plugging the above inequality back into (89) and (90), we further have

$$\lambda_1(\mathbf{W}) \le \hat{\lambda}_1 + \frac{1}{4}(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})) \le \hat{\lambda}_1 + \frac{1}{2}(\hat{\lambda}_1 - \hat{\lambda}_d), \tag{91}$$

$$\lambda_d(\mathbf{W}) \ge \hat{\lambda}_d - \frac{1}{4}(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})) \ge \hat{\lambda}_d - \frac{1}{2}(\hat{\lambda}_1 - \hat{\lambda}_d). \tag{92}$$

Let $\hat{\lambda}_{\max} = \max\{\hat{\lambda}_1, -\hat{\lambda}_d\}$. By (91) and (92), we can further bound the eigenvalues of $\mathbf{W}$ by

$$\lambda_1(\mathbf{W}) \le \hat{\lambda}_{\max} + \frac{1}{2} \cdot 2\hat{\lambda}_{\max} = 2\hat{\lambda}_{\max} \quad \text{and} \quad \lambda_d(\mathbf{W}) \ge -\hat{\lambda}_{\max} - \frac{1}{2} \cdot 2\hat{\lambda}_{\max} = -2\hat{\lambda}_{\max}. \tag{93}$$

Hence, we can see that $\|\mathbf{W}\|_{\text{op}} = \max\{\lambda_1(\mathbf{W}), -\lambda_d(\mathbf{W})\} \le 2\hat{\lambda}_{\max}$. Now we distinguish three cases.

(a) If $\hat{\lambda}_{\max} \le \frac{1}{2}$, then we are in **Case I** and the ExtEvec oracle outputs $\gamma = 2\hat{\lambda}_{\max} \le 1$ and $\mathbf{S} = 0$. In this case, we indeed have $\|\mathbf{W}\|_{\text{op}} \le \gamma \le 1$.

(b) If $\hat{\lambda}_{\max} \ge 2$, then we are in **Case II**. In addition, if $\hat{\lambda}_1 \ge -\hat{\lambda}_d$, then the ExtEvec oracle returns $\gamma = 2\hat{\lambda}_{\max}$ and $\mathbf{S} = 3\mathbf{u}^{(1)}(\mathbf{u}^{(1)})^\top$. Similarly, if $-\hat{\lambda}_d > \hat{\lambda}_1$, then the ExtEvec oracle returns $\gamma = 2\hat{\lambda}_{\max}$ and $\mathbf{S} = -3\mathbf{u}^{(d)}(\mathbf{u}^{(d)})^\top$. Without loss of generality, consider the case where $\hat{\lambda}_1 \ge -\hat{\lambda}_d$. Since $\|\mathbf{W}\|_{\text{op}} \le 2\hat{\lambda}_{\max} = \gamma$, we have $\|\mathbf{W}/\gamma\|_{\text{op}} \le 1$. Also, since $\mathbf{u}_1$ is a unit vector, we have $\|\mathbf{S}\|_F = 3\|\mathbf{u}^{(1)}\|^2 = 3$. Finally, for any $\hat{\mathbf{B}}$ such that $\|\hat{\mathbf{B}}\|_{\text{op}} \le 1$, we have

$$\langle \mathbf{S}, \mathbf{W} - \hat{\mathbf{B}} \rangle = 3(\mathbf{u}^{(1)})^\top \mathbf{W}\mathbf{u}^{(1)} - 3(\mathbf{u}^{(1)})^\top \hat{\mathbf{B}}\mathbf{u}^{(1)} \ge 3\hat{\lambda}_{\max} - 3 \ge 2\hat{\lambda}_{\max} - 1 = \gamma - 1,$$

where we used the fact that $\hat{\lambda}_{\max} \ge 2$ in the last inequality.

(c) If $\frac{1}{2} < \hat{\lambda}_{\max} < 2$, we continue to run the Lanczos method for a total number of $\left\lceil \frac{1}{4}\epsilon^{-1/2} \log \frac{11d}{q^2} + \frac{1}{2} \right\rceil$ iterations, where $\epsilon = \frac{1}{8}\delta$. Thus, by Proposition 3, with probability at least $1 - q/2$ we have

$$\tilde{\lambda}_1 \triangleq \langle \mathbf{W}\tilde{\mathbf{u}}^{(1)}, \tilde{\mathbf{u}}^{(1)} \rangle \ge \lambda_1(\mathbf{W}) - \frac{1}{8}\delta(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})), \tag{94}$$

$$\tilde{\lambda}_d \triangleq \langle \mathbf{W}\tilde{\mathbf{u}}^{(d)}, \tilde{\mathbf{u}}^{(d)} \rangle \le \lambda_d(\mathbf{W}) + \frac{1}{8}\delta(\lambda_1(\mathbf{W}) - \lambda_d(\mathbf{W})). \tag{95}$$

Let $\tilde{\lambda}_{\max} = \max\{\tilde{\lambda}_1, -\tilde{\lambda}_d\}$. Since we have $\lambda_1(\mathbf{W}) \le 2\hat{\lambda}_{\max} \le 4$ and $\lambda_d(\mathbf{W}) \ge -2\hat{\lambda}_{\max} \ge -4$, the above implies that $\tilde{\lambda}_1 \ge \lambda_1(\mathbf{W}) - \delta$ and $\tilde{\lambda}_d \le \lambda_d(\mathbf{W}) + \delta$. Hence, we can see that $\|\mathbf{W}\|_{\text{op}} = \max\{\lambda_1(\mathbf{W}), -\lambda_d(\mathbf{W})\} \le \hat{\lambda}_{\max} + \delta$. We further consider two subcases.

(c1) If $\tilde{\lambda}_{\max} \le 1 - \delta$, then we are in **Case I** and the ExtEvec oracle outputs $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = 0$. In this case, we indeed have $\|\mathbf{W}\|_{\text{op}} \le \gamma \le 1$.

(c2) If $\tilde{\lambda}_{\max} > 1 - \delta$, then we are in **Case II**. In addition, if $\tilde{\lambda}_1 \ge -\tilde{\lambda}_d$, then the ExtEvec oracle returns $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = \tilde{\mathbf{u}}^{(1)}(\tilde{\mathbf{u}}^{(1)})^\top$. Similarly, if $-\tilde{\lambda}_d > \tilde{\lambda}_1$, then the ExtEvec oracle returns $\gamma = \tilde{\lambda}_{\max} + \delta$ and $\mathbf{S} = -\tilde{\mathbf{u}}^{(d)}(\tilde{\mathbf{u}}^{(d)})^\top$. Without loss of generality, consider the case where $\tilde{\lambda}_1 \ge -\tilde{\lambda}_d$. Since $\|\mathbf{W}\|_{\text{op}} \le \tilde{\lambda}_{\max} + \delta = \gamma$, we have $\|\mathbf{W}/\gamma\|_{\text{op}} \le 1$. Also, since $\tilde{\mathbf{u}}^{(1)}$ is a unit vector, we have $\|\mathbf{S}\|_F = \|\tilde{\mathbf{u}}^{(1)}\|^2 = 1$. Finally, for any $\hat{\mathbf{B}}$ such that $\|\hat{\mathbf{B}}\|_{\text{op}} \le 1$, we have

$$\langle \mathbf{S}, \mathbf{W} - \hat{\mathbf{B}} \rangle = (\tilde{\mathbf{u}}^{(1)})^\top \mathbf{W}\tilde{\mathbf{u}}^{(1)} - (\tilde{\mathbf{u}}^{(1)})^\top \hat{\mathbf{B}}\tilde{\mathbf{u}}^{(1)} \ge \tilde{\lambda}_{\max} - 1 = \gamma - 1 - \delta.$$

This completes the proof. □

E.3 Proof of Theorem 2

We divide the proof of Theorem 2 into the following three lemmas.

Lemma 20. *If we run Algorithm 1 as specified in Theorem 1 for N iterations, then the total number of line search steps can be bounded by $2N + \log_{1/\beta}(\sigma_0 L_1/\alpha_2)$. As a corollary, the total number of gradient queries is bounded by $3N_\epsilon + \log_{1/\beta}(\frac{\sigma_0 L_1}{\alpha_2})$.*

Proof. In our backtracking scheme, the number of steps in each iteration is given by $\log_{1/\beta}(\eta_k/\hat{\eta}_k) + 1$. Also note that $\eta_{k+1} \leq \hat{\eta}_k/\beta$ for all $k \geq 0$. Thus, we have

$$\begin{aligned}\sum_{k=0}^{N-1}\left(\log_{1/\beta}\frac{\eta_k}{\hat{\eta}_k}+1\right) &= N + \log_{1/\beta}\frac{\sigma_0}{\hat{\eta}_0} + \sum_{k=0}^{N-2}\log_{1/\beta}\frac{\eta_{k+1}}{\hat{\eta}_{k+1}} \\ &\leq N + \log_{1/\beta}\frac{\sigma_0}{\hat{\eta}_0} + \sum_{k=0}^{N-2}\left(\log_{1/\beta}\frac{\hat{\eta}_k}{\hat{\eta}_{k+1}}+1\right) \\ &\leq 2N - 1 + \log_{1/\beta}\frac{\sigma_0}{\hat{\eta}_{N-1}}\end{aligned}$$

Furthermore, since $\hat{\eta}_k \geq \alpha_2\beta/L_1$ for all $k \geq 0$, we arrive at the conclusion. □

Lemma 21. *The total number of matrix-vector product evaluations in the* LinearSolver *oracle is bounded by $N_\epsilon + C_{11}\sqrt{\sigma_0 L_1} + C_{12}\sqrt{\frac{L_1\|\mathbf{z}_0-\mathbf{x}^*\|^2}{2\epsilon}}$, where C_{11} and C_{12} are absolute constants.*

Proof. The following proof loosely follows the strategy in [31]. We first bound the number of steps required by Subroutine 3 before it terminates.

Lemma 22. *Suppose $\mathbf{A} \succeq \mathbf{I}$. Then Subroutine 3 terminates after at most $\left\lceil\sqrt{\frac{\alpha+1}{\alpha}\lambda_{\max}(\mathbf{A})}-1\right\rceil$ iterations.*

Proof. Note that $\|\mathbf{s}_k\|_2 \geq \|\mathbf{s}^*\|_2 - \|\mathbf{s}_k - \mathbf{s}^*\|_2$. Also, since $\mathbf{A} \succeq \mathbf{I}$, we have $\|\mathbf{s}_k - \mathbf{s}^*\|_2 \leq \|\mathbf{A}(\mathbf{s}_k - \mathbf{s}^*)\|_2 = \|\mathbf{r}_k\|_2$. Therefore, we have

$$\|\mathbf{r}_k\|_2 \leq \alpha\|\mathbf{s}_k\|_2 \quad \Leftarrow \quad \|\mathbf{r}_k\|_2 \leq \alpha\|\mathbf{s}^*\|_2 - \alpha\|\mathbf{r}_k\|_2 \quad \Leftarrow \quad \|\mathbf{r}_k\|_2 \leq \frac{\alpha}{\alpha+1}\|\mathbf{s}^*\|_2.$$

By using Lemma 18, we only need $k \geq \sqrt{\frac{\alpha+1}{\alpha}\lambda_{\max}(\mathbf{A})} - 1$ to achieve $\|\mathbf{A}\mathbf{s}_k - \mathbf{b}\| \leq \alpha\|\mathbf{s}_k\|$. □

Moreover, when the step size is smaller enough, we can show that Subroutine 3 will terminate in one iteration.

Lemma 23. *Let $\mathbf{A} = \mathbf{I} + \eta\mathbf{B}$. When $\eta \leq \frac{\alpha}{2L_1}$, Algorithm 3 terminates in one iteration.*

Proof. From the update rule of Subroutine 3, we can compute that $\mathbf{s}_1 = \frac{\mathbf{b}^\top\mathbf{A}\mathbf{b}}{\|\mathbf{A}\mathbf{b}\|_2^2}\mathbf{b}$, which implies

$$\|\mathbf{s}_1\| = \|\mathbf{b}\| \cdot \frac{\|\mathbf{A}^{1/2}\mathbf{b}\|^2}{(\mathbf{A}^{1/2}\mathbf{b})^\top\mathbf{A}(\mathbf{A}^{1/2}\mathbf{b})} \geq \frac{\|\mathbf{b}\|}{\lambda_{\max}(\mathbf{A})} \geq \frac{\|\mathbf{b}\|}{1+\eta L_1}.$$

On the other hand, we also have

$$\|\mathbf{r}_1\| \leq \|\mathbf{A}\mathbf{b} - \mathbf{b}\| = \eta\|\mathbf{B}\mathbf{b}\| \leq \eta L_1\|\mathbf{b}\|. \tag{96}$$

Moreover, when $\eta \leq \frac{\alpha}{2L_1}$, we have $\eta L_1 \leq \frac{\alpha}{1+\eta L_1}$, which implies that $\|\mathbf{r}_1\| \leq \alpha\|\mathbf{s}_1\|$. □

Now we upper bound the total number of matrix-vector products in Algorithm 1. When $\mathbf{A} = \mathbf{I} + \eta_+\mathbf{B}_k$ where $\eta_+ = \eta_k\beta^i$. We can store the vector $\mathbf{B}_k\mathbf{b}$ at the beginning and reuse it to compute $\mathbf{s}_1$ when the step size $\eta_+ < \frac{\alpha_1}{2L_1}$. And when $\beta^i\eta_k L_1 \geq \frac{\alpha_1}{2}$, it holds that

$$1 + \beta^i\eta_k L_1 \leq \frac{\alpha_1+2}{\alpha_1}\beta^i\eta_k L_1.$$

Thus, at the k-th iteration, the number of matrix-vector products can be bounded by

$$\begin{aligned}\mathsf{MV}_k &\le 1+\sum_{i\ge 0,\eta_k\beta^i\ge\frac{\alpha_1}{2L_1}}\sqrt{\frac{\alpha_1+1}{\alpha_1}(1+\eta_k\beta^i L_1)}\\&\le 1+\sum_{i\ge 0,\eta_k\beta^i\ge\frac{\alpha_1}{2L_1}}\frac{\alpha_1+2}{\alpha_1}\sqrt{\beta^i\eta_k L_1}\\&\le 1+\frac{\alpha_1+2}{\alpha_1}\frac{1}{1-\sqrt{\beta}}\sqrt{\eta_k L_1}.\end{aligned}$$

Furthermore, we can bound that

$$\sum_{k=0}^{N-1}\sqrt{\eta_k}\le\sqrt{\sigma_0}+\sum_{k=1}^{N-1}\sqrt{\eta_k}\le\sqrt{\sigma_0}+\frac{1}{\sqrt{\beta}}\sum_{k=0}^{N-2}\sqrt{\hat{\eta}_k}\le\sqrt{\sigma_0}+\frac{2(2-\sqrt{\beta})}{\sqrt{\beta}(1-\sqrt{\beta})}\sqrt{A_{N-1}} \tag{97}$$

Note that $\epsilon < f(x_{N-1})-f(\mathbf{x}^*)\le\frac{\|\mathbf{z}_0-\mathbf{x}^*\|^2}{2A_{N-1}}$. Hence, we have $A_{N-1}\le\frac{\|\mathbf{z}_0-\mathbf{x}^*\|^2}{2\epsilon}$. Thus, we can bound the total number of matrix-vector product evaluations by

$$\begin{aligned}\mathsf{MV}=\sum_{k=0}^{N_\epsilon-1}\mathsf{MV}_k &\le N_\epsilon+\frac{\alpha_1+2}{\alpha_1}\frac{1}{1-\sqrt{\beta}}\left(\sqrt{\sigma_0 L_1}+\frac{2(2-\sqrt{\beta})}{\sqrt{\beta}(1-\sqrt{\beta})}\sqrt{\frac{L_1\|\mathbf{z}_0-\mathbf{x}^*\|^2}{2\epsilon}}\right),\\&=N_\epsilon+C_{11}\sqrt{\sigma_0 L_1}+C_{12}\sqrt{\frac{L_1\|\mathbf{z}_0-\mathbf{x}^*\|^2}{2\epsilon}},\end{aligned}$$

where we define $C_{11}=\frac{\alpha_1+2}{\alpha_1}\frac{1}{1-\sqrt{\beta}}$ and $C_{12}=\frac{\alpha_1+2}{\alpha_1}\frac{1}{1-\sqrt{\beta}}\frac{2(2-\sqrt{\beta})}{\sqrt{\beta}(1-\sqrt{\beta})}$. □

Lemma 24. *The total number of matrix-vector product evaluations in the* SEP *oracle is bounded by* $\mathcal{O}\big(N_\epsilon^{1.25}(\log N_\epsilon)^{0.5}\log\big(\frac{\sqrt{d}N_\epsilon}{p}\big)\big)$.

Proof. Note that we have $N_t\le\left\lceil\frac{1}{4\sqrt{2\delta_t}}\log\frac{44d}{q_t^2}+\frac{1}{2}\right\rceil$ in Subroutine 4, where $\delta_t=1/(\sqrt{t+2}\log(t+2))$ and $q_t=p/(2.5(t+1)\log^2(t+1))$. Thus, we have

$$N=\sum_{t=0}^{T-1}N_t\le\sum_{t=0}^{T-1}\frac{(t+2)^{0.25}\log^{0.5}(t+2)}{2\sqrt{2}}\log\frac{2.5\sqrt{44d}(t+1)\log^2(t+1)}{p} \tag{98}$$

$$=\mathcal{O}\left(N_\epsilon^{1.25}\sqrt{\log N_\epsilon}\log\frac{\sqrt{d}N_\epsilon}{p}\right). \tag{99}$$

□

F Experiments

In our experiments, we consider the logistic regression problem. Below we provide more details about the data generation scheme as well as the implementation of Nesterov's accelerated gradient method, BFGS, and our proposed A-QPNE algorithm.

Dataset generation. The dataset consists of n data points $\{(\mathbf{a}_i,y_i)\}_{i=1}^n$, where $\mathbf{a}_i\in\mathbb{R}^d$ is the i-th feature vector and $\mathbf{y}_i\in\{-1,1\}$ is its corresponding label. The labels $\{y_i\}_{i=1}^n$ are generated by

$$y_i=\operatorname{sign}(\langle\mathbf{a}_i^*,\mathbf{x}^*\rangle),\quad i=1,2,\dots,n,$$

where $\mathbf{a}_i^*\in\mathbb{R}^{d-1}$ and $\mathbf{x}^*\in\mathbb{R}^{d-1}$ are the underlying true feature vector and the underlying true parameter, respectively. Moreover, each entry of $\mathbf{a}_i^*$ and $\mathbf{x}^*$ is drawn independently according to the standard normal distribution $\mathcal{N}(0,1)$. Note that the true feature vectors $\{\mathbf{a}_i^*\}_{i=1}^n$ are not given in our dataset; instead, we generate $\{\mathbf{a}_i\}_{i=1}^n$ by adding noises and appending an extra dimension to $\{\mathbf{a}_i^*\}_{i=1}^n$. Specifically, we let $\mathbf{a}_i=[\mathbf{a}_i^*+\mathbf{n}_i+\mathbf{1};1]^\top\in\mathbb{R}^d$, where $\mathbf{n}_i\sim\mathcal{N}(0,\sigma^2\mathbf{I})$ is the i.i.d. Gaussian noise vector and $\mathbf{1} \in \mathbb{R}^{d-1}$ denotes the all-one vector. In our experiment, we set $n = 2,000$, $d = 150$ and $\sigma = 0.8$.

NAG. We implemented a monotone variant of the Nesterov accelerated gradient method as described in [43, Section 10.7.4]. Moreover, we determine the step size using a backtracking line search scheme.

BFGS. We implemented the classical BFGS algorithm, where the step size is determined by the Moré–Thuente line search scheme.

A-QPNE (our method). We implemented our proposed A-QPNE method following the pseudocode in Algorithm 1, where the line search scheme is given in Subroutine 1 and the Hessian approximation update is given in Subroutine 2. Moreover, the implementations of the LinearSolver oracle and the SEP oracle are given by Subroutines 3 and 4, respectively.