6 Appendix: Omitted Technical Details

In this section, we provide proofs for all the claims made in the main body of the paper, additional supporting propositions and lemmas, and also omitted details of the implementation of our algorithm. The proofs are provided in the order in which the corresponding statement appears in the main body. In Section 6.1, we prove the properties of our problem. We again emphasize that these properties are crucial to achieving our goal of a scale-invariant algorithm for (P). Section 6.2 contains the proofs of all our results pertaining to the convergence analysis of Algorithm 1, with proofs of the growth rate of the scalar sequences $\{a_i\}, \{a_i^k\}$, and A_k grouped separately in Section 6.3, owing to their more technical nature.

6.1 Omitted Proof from Section 2: Properties of Our Objective

Proposition 2.1. *Given $f : \mathbb{R}^n_+ \to \mathbb{R}$ as defined in (2) and $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^n_+} f(\mathbf{x})$, the following statements all hold.*

a) $\nabla f(\mathbf{x}^\star) \geq \mathbf{0}$.
b) $f(\mathbf{x}^\star) = -\frac{1}{2}\|\mathbf{A}\mathbf{x}^\star\|_2^2 = -\frac{1}{2}\mathbf{1}^\top\mathbf{x}^\star$.
c) for all $j \in [n]$, we have $x_j^\star \in \left[0, \frac{1}{\|\mathbf{A}_{:j}\|_2^2}\right]$.
d) $-\frac{1}{2}\sum_{j\in[n]} \frac{1}{\|\mathbf{A}_{:j}\|_2^2} \leq f(\mathbf{x}^\star) \leq -\frac{1}{2\min_{j\in[n]}\|\mathbf{A}_{:j}\|_2^2}$.

Proof. We recall the first-order optimality condition stated in Inequality (1): for all $\mathbf{x} \geq \mathbf{0}$, we have $\langle \nabla f(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star\rangle \geq 0$; we repeatedly invoke this inequality in the proof below.

1. Suppose there exists a coordinate j at which Proposition 2.1 (a) does not hold and instead, we have $\nabla_j f(\mathbf{x}^\star) < 0$. Consider $\mathbf{x} \geq \mathbf{0}$ such that $x_i = x_i^\star$ for all $i \neq j$ and let $x_j = x_j^\star + \varepsilon$ for some $\varepsilon > 0$. Then, Inequality (1) becomes $\nabla_j f(\mathbf{x}^\star) \cdot \varepsilon \geq 0$. Under the assumption $\nabla_j f(\mathbf{x}^\star) < 0$, this is an invalid inequality, thus contradicting our assumption.

2. From Proposition 2.1 (a), we know that $\nabla f(\mathbf{x}^\star) \geq \mathbf{0}$. If $\nabla_i f(\mathbf{x}^\star) > 0$, and if $x_i^\star > 0$, then by picking a vector $\mathbf{x}$ such that $x_j = x_j^\star$ for $j \neq i$ and $x_i = x_i^\star - \gamma$ for any $\gamma \in (0, x_i^\star)$, we violate Inequality (1). Therefore it must be the case that if $\nabla_i f(\mathbf{x}^\star) > 0$, then $x_i^\star = 0$. Thus we have
$$0 = \langle \mathbf{x}^\star, \nabla f(\mathbf{x}^\star)\rangle = \langle \mathbf{x}^\star, \mathbf{A}^\top\mathbf{A}\mathbf{x}^\star - \mathbf{1}\rangle.$$
Therefore, $f(\mathbf{x}^\star) = \frac{1}{2}\langle \mathbf{x}^\star, \mathbf{A}^\top\mathbf{A}\mathbf{x}^\star\rangle - \mathbf{1}^\top\mathbf{x}^\star = -\frac{1}{2}\langle \mathbf{x}^\star, \mathbf{A}^\top\mathbf{A}\mathbf{x}^\star\rangle = -\frac{1}{2}\mathbf{1}^\top\mathbf{x}^\star$.

3. From the proof of Proposition 2.1 (b), we have $\langle \mathbf{x}^\star, \nabla f(\mathbf{x}^\star)\rangle = 0$. We also have $\mathbf{x}^\star \geq \mathbf{0}$ and, from Proposition 2.1 (a), that $\nabla f(\mathbf{x}^\star) \geq \mathbf{0}$. Therefore, if $\mathbf{x}_i^\star > 0$ for some coordinate i then it must be that $\nabla_i f(\mathbf{x}^\star) = 0$. That is, $1 = \langle \mathbf{A}_{:i}, \mathbf{A}\mathbf{x}^\star\rangle$. Combining this equality with the fact that $\mathbf{A}$ and $\mathbf{x}^\star$ are both coordinate-wise non-negative gives
$$1 = \langle \mathbf{A}_{:i}, \mathbf{A}\mathbf{x}^\star\rangle \geq \langle \mathbf{A}_{:i}, \mathbf{A}_{:i}x_i^\star\rangle,$$
which implies $x_i^\star \leq \frac{1}{\|\mathbf{A}_{:i}\|_2^2}$ for all coordinates i.

4. The lower bound follows immediately by combining Proposition 2.1 (b) and Proposition 2.1 (c). For the upper bound, we need to find *a* feasible point $\widehat{\mathbf{x}}$ and compute the function value at $\widehat{\mathbf{x}}$, since $f(\mathbf{x}^\star) = \min_{\mathbf{y}\geq\mathbf{0}} f(\mathbf{y}) \leq f(\widehat{\mathbf{x}})$. Let $\widehat{\mathbf{x}} = \gamma\mathbf{e}_k$ for some $\gamma > 0$. Then,
$$f(\widehat{\mathbf{x}}) = \frac{1}{2}\gamma^2\|\mathbf{A}_{:k}\|_2^2 - \gamma.$$
Let $\gamma = \frac{1}{\|\mathbf{A}_{:k}\|_2^2}$. Then, $f(\widehat{\mathbf{x}}) = -\frac{1}{2\|\mathbf{A}_{:k}\|_2^2}$. We pick $k = \arg\min_{i\in[n]}\|\mathbf{A}_{:i}\|_2$, therefore $f(\mathbf{x}^\star) \leq -\frac{1}{2\min_{i\in[n]}\|\mathbf{A}_{:i}\|_2^2}$ as claimed.

□

6.2 Omitted Proofs from Section 3: Analysis of Algorithm

6.2.1 Proofs from Section 3.1: Results on Upper and Lower Estimates

We first show the results stating U_k and L_k are indeed valid upper and lower (respectively) estimates of the Lagrangian.

Lemma 3.1. *For U_k as defined in Eq. (10), Lagrangian defined in Eq. (3) and $\widetilde{\mathbf{x}}_k \in \mathbb{R}^n_+$ in Eq. (8), we have, for all $\mathbf{y} \in \mathbb{R}^m$, the upper bound $\mathrm{U}_k(\mathbf{y}) \geq \mathcal{L}(\widetilde{\mathbf{x}}_k, \mathbf{y})$.*

Proof. By evaluating the Lagrangian described by Eq. (3) at $\mathbf{x} = \widetilde{\mathbf{x}}_k$, and by definition of ψ_k from Eq. (9), we obtain the following upper bound on the Lagrangian at $(\widetilde{\mathbf{x}}_k, \mathbf{y})$.

$$\begin{aligned}
\mathcal{L}(\widetilde{\mathbf{x}}_k, \mathbf{y}) &= \langle \mathbf{A}\widetilde{\mathbf{x}}_k, \mathbf{y}\rangle - \frac{1}{2}\|\mathbf{y}\|_2^2 - \mathbf{1}^\top \widetilde{\mathbf{x}}_k \\
&= \frac{1}{A_k}\sum_{i\in[k]} a_i^k \left[\langle \mathbf{A}\mathbf{x}_i, \mathbf{y}\rangle - \frac{1}{2}\|\mathbf{y}\|^2 - \mathbf{1}^\top \mathbf{x}_i\right] = \frac{1}{A_k}\psi_k(\mathbf{y}) \\
&\leq \frac{1}{A_k}\psi_k(\mathbf{y}_k) - \frac{1}{2}\|\mathbf{y} - \mathbf{y}_k\|_2^2 = \mathrm{U}_k(\mathbf{y}),
\end{aligned}$$

where the final steps are by strong convexity of ψ_k and by Eq. (10). □

We emphasize that $\mathbf{y}$ here can be random since this is a deterministic statement.

Lemma 3.2. *For L_k defined in Eq. (17), for the Lagrangian in Eq. (3) and $\widetilde{\mathbf{y}}_k$ in Eq. (8), we have, for a fixed $\mathbf{u} \in \mathcal{X}$, the lower bound $\mathbb{E}\mathcal{L}(\mathbf{u}, \widetilde{\mathbf{y}}_k) \geq \mathbb{E}\mathrm{L}_k(\mathbf{u})$, where the expectation is with respect to all the random choices of coordinates in Algorithm 1.*

Proof. First, evaluating Eq. (3) at $\widetilde{\mathbf{y}}_k$ gives

$$\mathcal{L}(\mathbf{u}, \widetilde{\mathbf{y}}_k) = \langle \mathbf{A}\mathbf{u}, \widetilde{\mathbf{y}}_k\rangle - \mathbf{1}^\top \mathbf{u} - \frac{1}{2}\|\widetilde{\mathbf{y}}_k\|_2^2.$$

Taking the expectation on both sides, applying the definition of $\widetilde{\mathbf{y}}_k$, convexity of $\frac{1}{2}\|\cdot\|^2$, and Jensen's inequality, and adding and subtracting $\frac{1}{A_k}\mathbb{E}\sum_{i\in[k]} a_i\langle \mathbf{A}\mathbf{x}, \overline{\mathbf{y}}_{i-1}\rangle + \frac{1}{A_k}\phi_0(\mathbf{u})$ gives

$$\begin{aligned}
\mathbb{E}\mathcal{L}(\mathbf{u}, \widetilde{\mathbf{y}}_k) &\geq \frac{1}{A_k}\mathbb{E}\left[\sum_{i\in[k]} a_i\left[\langle \mathbf{A}\mathbf{u}, \mathbf{y}_i\rangle - \mathbf{1}^\top\mathbf{u} - \frac{1}{2}\|\mathbf{y}_i\|_2^2\right]\right] \\
&= \frac{1}{A_k}\mathbb{E}\left[\sum_{i\in[k]} a_i\left[\langle \mathbf{A}\mathbf{u}, \overline{\mathbf{y}}_{i-1}\rangle - \mathbf{1}^\top\mathbf{u} - \frac{1}{2}\|\mathbf{y}_i\|_2^2\right]\right] + \frac{1}{A_k}\mathbb{E}\left[\phi_0(\mathbf{u})\right] - \frac{1}{A_k}\mathbb{E}\left[\phi_0(\mathbf{u})\right] \\
&\quad + \frac{1}{A_k}\mathbb{E}\left[\sum_{i\in[k]} a_i\langle \mathbf{A}\mathbf{u}, \mathbf{y}_i - \overline{\mathbf{y}}_{i-1}\rangle\right].
\end{aligned}$$

We continue the analysis as

$$\begin{aligned}
\mathbb{E}\mathcal{L}(\mathbf{u}, \widetilde{\mathbf{y}}_k) &\geq \frac{1}{A_k}\mathbb{E}\left[\phi_k(\mathbf{u})\right] - \frac{1}{A_k}\mathbb{E}\left[\phi_0(\mathbf{u})\right] + \frac{1}{A_k}\sum_{i\in[k]} a_i\mathbb{E}\langle \mathbf{A}\mathbf{u}, \mathbf{y}_i - \overline{\mathbf{y}}_{i-1}\rangle - \frac{1}{2A_k}\mathbb{E}\sum_{i\in[k]} a_i\|\mathbf{y}_i\|_2^2 \\
&\geq \frac{1}{A_k}\mathbb{E}\left[\phi_k(\mathbf{x}_k)\right] + \mathbb{E}\left[\frac{1}{2A_k}\|\mathbf{u} - \mathbf{x}_k\|_{\mathbf{\Lambda}}^2\right] - \frac{1}{A_k}\mathbb{E}[\phi_0(\mathbf{u})] + \frac{1}{A_k}\mathbb{E}\left[\sum_{i\in[k]} a_i\langle \mathbf{A}\mathbf{u}, \mathbf{y}_i - \overline{\mathbf{y}}_{i-1}\rangle\right] \\
&\quad - \frac{1}{2A_k}\mathbb{E}\sum_{i\in[k]} a_i\|\mathbf{y}_i\|_2^2 \\
&= \mathbb{E}\mathrm{L}_k(\mathbf{u}),
\end{aligned}$$

the first step comes from Eq. (13), the second step comes from Eq. (16), and the final step comes from Eq. (17). □

6.2.2 Proofs from Section 3.2

We now describe three technical propositions that bound terms that show up in the proof of our result on bounding the scaled gap estimate.

Proposition 6.1. *For ψ_k defined in Eq. (9), with $\{a_i^k\}$ defined in Eq. (7), we have for all $k \geq 1$,*

$$\begin{aligned}\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1}) \leq a_k^k \left\{ \langle \mathbf{y}_k, \mathbf{A}\mathbf{x}_k \rangle - \frac{1}{2}\|\mathbf{y}_k\|_2^2 - \mathbf{1}^\top \mathbf{x}_k \right\} \\ + \sum_{i=1}^{k-1} (a_i^k - a_i^{k-1}) \left[\langle \mathbf{y}_k, \mathbf{A}\mathbf{x}_i \rangle - \frac{1}{2}\|\mathbf{y}_k\|_2^2 - \mathbf{1}^\top \mathbf{x}_i \right] \\ - \frac{A_{k-1}}{2}\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2.\end{aligned}$$

Proof. Evaluating ψ_k and ψ_{k-1} as defined in Eq. (9) at $\mathbf{y}_k$ and subtracting, we have

$$\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_k) = a_k \langle \mathbf{A}^\top \mathbf{y}_k - \mathbf{1}, n\mathbf{x}_k - (n-1)\mathbf{x}_{k-1} \rangle - \frac{a_k}{2}\|\mathbf{y}_k\|_2^2. \tag{20}$$

Next, applying strong convexity of ψ_{k-1} at $\mathbf{y}_k$ and $\mathbf{y}_{k-1}$ while using the fact that $\mathbf{y}_{k-1}$ minimizes ψ_{k-1} gives

$$\psi_{k-1}(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1}) \leq -\frac{1}{2} A_{k-1}\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2. \tag{21}$$

To complete the proof, it remains to add Eq. (20) and Inequality (21). □

Proposition 6.2. *The random function $\phi_k : \mathcal{X} \to \mathbb{R}$, $k \geq 2$, defined in Eq. (15) satisfies the following properties, with $\mathbf{x}_k$, $\mathbf{y}_k$, and $\overline{\mathbf{y}}_k$ evolving as per Algorithm 1.*

a) It is separable in its coordinates: $\phi_k(\mathbf{x}) = \sum_{j\in[n]} \phi_{k,j}(x_j)$, where, for each $j \in [n]$, we define $\phi_{0,j}(x_j) = \frac{\|\mathbf{A}_{:j}\|_2^2}{2}(x_j - [\mathbf{x}_0]_j)^2$, $\phi_{1,j}(x_j) = a_1 x_j (\mathbf{A}^\top \overline{\mathbf{y}}_0 - \mathbf{1})_j + \phi_{0,j}(x_j)$, and for $k \geq 2$,

$$\phi_{k,j}(x_j) = \phi_{k-1,j}(x_j) + n a_k \mathbf{1}_{j=j_k} \langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, x_{j_k} \mathbf{e}_{j_k} \rangle. \tag{22}$$

b) The primal variable $\mathbf{x}_k$ is updated only on the j_k^{th} coordinate in each iteration: $\mathbf{x}_k = \mathbf{x}_{k-1} + \gamma \mathbf{e}_{j_k}$ for some γ, and $[\mathbf{x}_k]_j = [\mathbf{x}_{k-1}]_j$ for $j \neq j_k$.

c) For a fixed $\mathbf{x} \in \mathcal{X}$ and for $k \geq 1$, we have, over all the randomness in the algorithm,

$$\mathbb{E}\left[\phi_k(\mathbf{x})\right] = \mathbb{E}\left[\phi_0(\mathbf{x})\right] + \sum_{i\in[k]} a_i \mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{i-1} - \mathbf{1}, \mathbf{x} \rangle\right]. \tag{23}$$

Proof. In the statement of Proposition 6.2 (a), the claim about separability of ϕ_0 and ϕ_1 can be checked just from the definitions of $\phi_{0,j}$ and $\phi_{1,j}$. We prove the claim of coordinate-wise separability for $k \geq 2$ by summing over $j \in [n]$ both sides of Eq. (22). We can compute this sum via following observation, which concludes the proof of Proposition 6.2 (a).

$$\sum_{j\in[n]} a_k \mathbf{1}_{j=j_k} \langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, x_{j_k} \mathbf{e}_{j_k} \rangle = a_k \langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, x_{j_k} \mathbf{e}_{j_k} \rangle.$$

From Proposition 6.2 (a), we may therefore define, for $j \neq j_k$,

$$[\mathbf{x}_k]_j = \arg\min_{u\in\mathbb{R}_+} \phi_{k,j}(u) = \arg\min_{u\in\mathbb{R}_+} \phi_{k-1,j}(u) = [\mathbf{x}_{k-1}]_j.$$

Therefore, $[\mathbf{x}_k]_j = [\mathbf{x}_{k-1}]_j$ for all $j \neq j_k$, thus proving Proposition 6.2 (b). To prove Proposition 6.2 (c), we use induction on k. The base case holds for $k = 1$ by the definition of $\phi_1(\mathbf{x})$. Let Proposition 6.2 (c) hold for $k \geq 1$. Then, by the definition of ϕ_k as in Eq. (15), we have

$$\phi_k(\mathbf{x}) = \phi_{k-1}(\mathbf{x}) + n a_k \langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, x_{j_k} \mathbf{e}_{j_k} \rangle, \text{ for all } k \geq 2.$$

Let $\mathcal{F}_{k-1}$ be the natural filtration, containing all the randomness in the algorithm up to and including iteration $k-1$. Taking expectations with respect to all the randomness until iteration k and invoking linearity of expectation, the inductive hypothesis, and the tower rule $\mathbb{E}[\,\cdot\,] = \mathbb{E}[\mathbb{E}[\,\cdot\,|\mathcal{F}_{k-1}]]$, we have

$$\begin{aligned}\mathbb{E}[\phi_k(\mathbf{x})] &= \mathbb{E}[\phi_{k-1}(\mathbf{x})] + na_k\mathbb{E}\left[\left[\mathbb{E}\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, x_{j_k}\mathbf{e}_{j_k}\rangle|\mathcal{F}_{k-1}\right]\right] \\ &= \mathbb{E}\left[\phi_0(\mathbf{x})\right] + \sum_{i\in[k-1]} a_i\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{i-1} - \mathbf{1}, \mathbf{x}\rangle\right] + a_k\mathbb{E}\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, \mathbf{x}\rangle \\ &= \mathbb{E}\left[\phi_0(\mathbf{x})\right] + \sum_{i\in[k]} a_i\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{i-1} - \mathbf{1}, \mathbf{x}\rangle\right],\end{aligned}$$

which finishes the proof of Proposition 6.2 (c). □

Proposition 6.3. *For all $k \geq 2$, the random function $\phi_k : \mathcal{X} \to \mathbb{R}$, $k \geq 2$, defined in Eq. (15) satisfies the following inequality, where $\mathbf{x}_k$ and $\overline{\mathbf{y}}_k$ evolve according to Algorithm 1.*

$$\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1}) \geq a_k\left(n\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, [\mathbf{x}_k]_{j_k}\mathbf{e}_{j_k}\rangle\right) + \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2.$$

Proof. We have, using $\mathbf{x} = \mathbf{x}_k$ in Eq. (15), that

$$\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_k) = na_k\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, [\mathbf{x}_k]_{j_k}\mathbf{e}_{j_k}\rangle.$$

Applying Eq. (16) to ϕ_{k-1} at $\mathbf{x}_k$ gives

$$\phi_{k-1}(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1}) \geq \frac{1}{2}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2.$$

Adding these inequalities completes the proof of the claim. □

We now use the preceding technical results to bound the change in scaled gap.

Lemma 3.3. *Consider the iterates $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ evolving according to Algorithm 1. Let $n \geq 2$ and assume that $a_1 = \frac{1}{\sqrt{2}n^{1.5}}$ and $a_1 \geq (n-1)a_2$, while for $k \geq 3$,*

$$a_k \leq \min\left(\frac{na_{k-1}}{n-1}, \frac{\sqrt{A_{k-1}}}{2n}\right). \tag{18}$$

Then, for fixed $\mathbf{u} \in \mathcal{X}$, any $\mathbf{v} \in \mathbb{R}^m$, and all $k \geq 2$, the gap estimate $\mathrm{G}_k = \mathrm{U}_k - \mathrm{L}_k$ satisfies

$$\begin{aligned}&\mathbb{E}(A_kG_k(\mathbf{x},\mathbf{y}) - A_{k-1}G_{k-1}(\mathbf{x},\mathbf{y})) \\ &\leq -\mathbb{E}\left(\frac{A_k}{2}\|\mathbf{y} - \mathbf{y}_k\|_2^2 - \frac{A_{k-1}}{2}\|\mathbf{y} - \mathbf{y}_{k-1}\|_2^2\right) - \frac{1}{2}\mathbb{E}\|\mathbf{x} - \mathbf{x}_k\|_{\mathbf{\Lambda}}^2 + \frac{1}{2}\mathbb{E}\|\mathbf{x} - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2 \\ &\quad - a_k\mathbb{E}\langle \mathbf{A}(\mathbf{x} - \mathbf{x}_k), \mathbf{y}_k - \mathbf{y}_{k-1}\rangle + a_{k-1}\mathbb{E}\langle \mathbf{A}(\mathbf{x} - \mathbf{x}_{k-1}), \mathbf{y}_{k-1} - \mathbf{y}_{k-2}\rangle \\ &\quad - \frac{1}{4}A_{k-1}\mathbb{E}\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2 + \frac{1}{4}A_{k-2}\mathbb{E}\|\mathbf{y}_{k-1} - \mathbf{y}_{k-2}\|_2^2.\end{aligned}$$

Proof. Using $\mathrm{G}_k = \mathrm{U}_k - \mathrm{L}_k$, U_k from Eq. (10), and L_k from Eq. (17), we have

$$\begin{aligned}A_k\mathrm{G}_k(\mathbf{u},\mathbf{v}) &= \psi_k(\mathbf{y}_k) - \phi_k(\mathbf{x}_k) + \phi_0(\mathbf{u}) \\ &\quad - \frac{A_k}{2}\|\mathbf{v} - \mathbf{y}_k\|_2^2 - \frac{1}{2}\|\mathbf{u} - \mathbf{x}_k\|_{\mathbf{\Lambda}}^2 - \sum_{i\in[k]} a_i\langle \mathbf{A}\mathbf{u}, \mathbf{y}_i - \overline{\mathbf{y}}_{i-1}\rangle + \sum_{i\in[k]}\frac{a_i}{2}\|\mathbf{y}_i\|_2^2.\end{aligned}$$

Therefore, the difference in scaled gap between successive iterations is

$$\begin{aligned}A_k\mathrm{G}_k(\mathbf{u},\mathbf{v}) - A_{k-1}\mathrm{G}_{k-1}(\mathbf{u},\mathbf{v}) = \;& [\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1})] - [\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1})] \\ &- \frac{A_k}{2}\|\mathbf{v} - \mathbf{y}_k\|_2^2 + \frac{A_{k-1}}{2}\|\mathbf{v} - \mathbf{y}_{k-1}\|_2^2 \\ &- \frac{1}{2}\|\mathbf{u} - \mathbf{x}_k\|_{\mathbf{\Lambda}}^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2 \\ &- a_k\langle \mathbf{A}\mathbf{u}, \mathbf{y}_k - \overline{\mathbf{y}}_{k-1}\rangle + \frac{a_k}{2}\|\mathbf{y}_k\|_2^2.\end{aligned} \tag{24}$$

Based on the above expression, to prove the lemma, it suffices to bound the expectation of

$$T_k(\mathbf{u}) \stackrel{\text{def}}{=} [\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1})] - [\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1})] - a_k\langle \mathbf{A}\mathbf{u}, \mathbf{y}_k - \overline{\mathbf{y}}_{k-1}\rangle + \frac{a_k}{2}\|\mathbf{y}_k\|_2^2. \tag{25}$$

First, we take expectations on both sides of the inequality in Proposition 6.3 by invoking $\mathbb{E}[\,\cdot\,] = \mathbb{E}[\mathbb{E}[\,\cdot\,|\mathcal{F}_{k-1}]]$ as before, where $\mathcal{F}_{k-1}$ denotes the natural filtration. By using the fact that $\mathbf{x}_{k-1}$ is updated only at coordinate j_k (as stated in Proposition 6.2 (b)), we observe the following for the term from the right hand side of Proposition 6.3.

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, [\mathbf{x}_k]_{j_k}\mathbf{e}_{j_k}\rangle\right] \\
=&\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, \mathbf{x}_k - \mathbf{x}_{k-1}\rangle\right] + \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, [\mathbf{x}_{k-1}]_{j_k}\mathbf{e}_{j_k}\rangle|\mathcal{F}_{k-1}\right]\right] \\
=&\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, \mathbf{x}_k - \mathbf{x}_{k-1}\rangle\right] + \frac{1}{n}\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, \mathbf{x}_{k-1}\rangle\right] \\
=&\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, \mathbf{x}_k - (1-1/n)\,\mathbf{x}_{k-1}\rangle\right].
\end{aligned} \tag{26}$$

Therefore, we have from Proposition 6.3 and scaling Eq. (26) by $-na_k$ that

$$\begin{aligned}
-\mathbb{E}\left[\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1})\right] \leq\ & -\frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2\right] \\
& - a_k\mathbb{E}\left[\langle \mathbf{A}^\top \overline{\mathbf{y}}_{k-1} - \mathbf{1}, n\mathbf{x}_k - (n-1)\mathbf{x}_{k-1}\rangle\right].
\end{aligned} \tag{27}$$

We now bound the expectation of the term involving differences of ψ_k by taking expectations of both sides of Proposition 6.1.

$$\begin{aligned}
\mathbb{E}\left[\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1})\right] \leq & -\frac{A_{k-1}}{2}\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2\right] - \frac{a_k}{2}\mathbb{E}\left[\|\mathbf{y}_k\|_2^2\right] \\
& + a_k\mathbb{E}\left[\langle \mathbf{A}^\top \mathbf{y}_k - \mathbf{1}, n\mathbf{x}_k - (n-1)\mathbf{x}_{k-1}\rangle\right].
\end{aligned} \tag{28}$$

By taking expectations on both sides of Eq. (25), we have

$$\begin{aligned}
\mathbb{E}[T_k(\mathbf{u})] = & \mathbb{E}\left[\psi_k(\mathbf{y}_k) - \psi_{k-1}(\mathbf{y}_{k-1})\right] - \mathbb{E}\left[\phi_k(\mathbf{x}_k) - \phi_{k-1}(\mathbf{x}_{k-1})\right] \\
& - a_k\mathbb{E}\left[\langle \mathbf{A}\mathbf{u}, \mathbf{y}_k - \overline{\mathbf{y}}_{k-1}\rangle\right] + \frac{a_k}{2}\mathbb{E}\left[\|\mathbf{y}_k\|_2^2\right].
\end{aligned} \tag{29}$$

Combining Eq. (27), Eq. (28), and Eq. (29) then gives

$$\begin{aligned}
\mathbb{E}[T_k(\mathbf{u})] \leq & -\frac{A_{k-1}}{2}\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2\right] \\
& + a_k\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_k - \overline{\mathbf{y}}_{k-1}), n\mathbf{x}_k - (n-1)\mathbf{x}_{k-1} - \mathbf{u}\rangle\right].
\end{aligned} \tag{30}$$

Recall that by the assumption in the statement of the lemma,

$$\overline{\mathbf{y}}_{k-1} = \mathbf{y}_{k-1} + \frac{a_{k-1}}{a_k}(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}).$$

Plugging into Eq. (30) and rearranging, we have

$$\begin{aligned}
\mathbb{E}[T_k(\mathbf{u})] \leq & -\frac{A_{k-1}}{2}\mathbb{E}\left[\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2\right] \\
& + (n-1)a_k\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_k - \mathbf{y}_{k-1}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle\right] - na_{k-1}\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle\right] \\
& + a_k\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_k - \mathbf{y}_{k-1}), \mathbf{x}_k - \mathbf{u}\rangle\right] - a_{k-1}\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \mathbf{x}_{k-1} - \mathbf{u}\rangle\right].
\end{aligned} \tag{31}$$

To complete the proof, we need to bound the terms from the first two lines on the right-hand side of Eq. (31). First, observe that, by the coordinate update of $\mathbf{x}_k$ and Young's inequality, $\forall \beta > 0$,

$$\begin{aligned}
\langle \mathbf{A}^\top(\mathbf{y}_k - \mathbf{y}_{k-1}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle &= \langle \mathbf{y}_k - \mathbf{y}_{k-1}, \mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1})\rangle \\
&= \langle \mathbf{y}_k - \mathbf{y}_{k-1}, \mathbf{A}_{:j_k}([\mathbf{x}_k]_{j_k} - [\mathbf{x}_{k-1}]_{j_k})\rangle \\
&\leq \frac{\beta}{2}\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2 + \frac{1}{2\beta}\|\mathbf{A}_{:j_k}\|_2^2|[\mathbf{x}_k]_{j_k} - [\mathbf{x}_{k-1}]_{j_k}|^2 \\
&= \frac{\beta}{2}\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2 + \frac{1}{2\beta}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2.
\end{aligned} \tag{32}$$

By the same token, $\forall \gamma > 0$,

$$-\langle \mathbf{A}^\top(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \mathbf{x}_k - \mathbf{x}_{k-1}\rangle \leq \frac{\gamma}{2}\|\mathbf{y}_{k-1} - \mathbf{y}_{k-2}\|_2^2 + \frac{1}{2\gamma}\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_{\mathbf{\Lambda}}^2. \tag{33}$$

Recalling that, by the choice of step sizes, $(n-1)a_k \leq n a_{k-1}$ and $a_k \leq \frac{\sqrt{A_{k-1}}}{2n}$, we can verify that for $\beta = 2(n-1)a_k$ and $\gamma = 2n a_{k-1}$, the following inequalities hold:

$$\begin{aligned} (n-1)a_k\beta - A_{k-1} &\leq -\frac{A_{k-1}}{2}, \\ \frac{(n-1)a_k}{\beta} + \frac{n a_{k-1}}{\gamma} &\leq 1. \end{aligned} \tag{34}$$

Combining Equations (31)–(34),

$$\begin{aligned} \mathbb{E}[T_k(\mathbf{u})] \leq & -\frac{A_{k-1}}{4}\mathbb{E}[\|\mathbf{y}_k - \mathbf{y}_{k-1}\|_2^2] + n^2 {a_{k-1}}^2 \mathbb{E}[\|\mathbf{y}_{k-1} - \mathbf{y}_{k-2}\|_2^2] \\ & + a_k \mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_k - \mathbf{y}_{k-1}), \mathbf{x}_k - \mathbf{u}\rangle\right] - a_{k-1}\mathbb{E}\left[\langle \mathbf{A}^\top(\mathbf{y}_{k-1} - \mathbf{y}_{k-2}), \mathbf{x}_{k-1} - \mathbf{u}\rangle\right]. \end{aligned} \tag{35}$$

It remains to combine Eq. (24), Eq. (25), and Eq. (35). □

Lemma 3.4. *Given a fixed $\mathbf{u} \in \mathcal{X}$, any $\mathbf{v} \in \mathbb{R}^m$, $\bar{\mathbf{y}}_0 = \mathbf{y}_0$, and $\mathbf{x}_1$ and $\mathbf{y}_1$ from Algorithm 1, we have*

$$A_1 \mathrm{G}_1(\mathbf{u}, \mathbf{v}) = a_1\langle \mathbf{A}^\top(\mathbf{y}_1 - \mathbf{y}_0), \mathbf{x}_1 - \mathbf{u}\rangle + \phi_0(\mathbf{u}) - \phi_0(\mathbf{x}_1) - \frac{1}{2}\|\mathbf{u} - \mathbf{x}_1\|_{\mathbf{\Lambda}}^2 - \frac{A_1}{2}\|\mathbf{v} - \mathbf{y}_1\|_2^2.$$

Proof. Evaluating Eq. (10) and Eq. (17) at $k = 1$ gives

$$A_1 U_1(\mathbf{v}) = \psi_1(\mathbf{y}_1) - \frac{A_1}{2}\|\mathbf{v} - \mathbf{y}_1\|_2^2, \tag{36}$$

$$A_1 L_1(\mathbf{u}) = \phi_1(\mathbf{x}_1) + \frac{1}{2}\|\mathbf{u} - \mathbf{x}_1\|_{\mathbf{\Lambda}}^2 - \phi_0(\mathbf{u}) + a_1\langle \mathbf{A}\mathbf{u}, \mathbf{y}_1 - \overline{\mathbf{y}}_0\rangle - \frac{a_1}{2}\|\mathbf{y}_1\|_2^2. \tag{37}$$

By definition of ψ_1 from Eq. (9), ϕ_1 from Eq. (12), and the assignment $a_1^1 = a_1$, we have

$$\begin{aligned} \psi_1(\mathbf{y}_1) - \phi_1(\mathbf{x}_1) &= -\frac{a_1}{2}\|\mathbf{y}_1\|_2^2 + a_1\langle \mathbf{y}_1, \mathbf{A}\mathbf{x}_1\rangle - a_1 \mathbf{1}^\top \mathbf{x}_1 - \phi_0(\mathbf{x}_1) - a_1\langle \mathbf{A}^\top \overline{\mathbf{y}}_0 - \mathbf{1}, \mathbf{x}_1\rangle \\ &= -\frac{a_1}{2}\|\mathbf{y}_1\|_2^2 + a_1\langle \mathbf{A}(\mathbf{y}_1 - \mathbf{y}_0), \mathbf{x}_1\rangle - \phi_0(\mathbf{x}_1), \end{aligned}$$

where we have used that $\bar{\mathbf{y}}_0 = \mathbf{y}_0$, which holds by assumption. To complete the proof, it remains to subtract Eq. (37) from Eq. (36) and combine with the last equality. □

Theorem 3.5. *[Main Result] Assume that $n \geq 4$. Given a matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, $\varepsilon > 0$, an arbitrary $\mathbf{x}_0 \in \mathcal{X}$ and $\bar{\mathbf{y}}_0 = \mathbf{y}_0 = \mathbf{A}\mathbf{x}_0$, let $\mathbf{x}_k$ and A_k evolve according to SI-NNLS+ (Algorithm 1) for $k \geq 1$. For f defined in (2), define $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x} \geq \mathbf{0}} f(\mathbf{x})$. Then, for all $K \geq 2$, we have*

$$\mathbb{E}\Big[\langle \nabla f(\widetilde{\mathbf{x}}_K), \widetilde{\mathbf{x}}_K - \mathbf{x}^\star\rangle + \frac{1}{2}\|\mathbf{A}(\widetilde{\mathbf{x}}_K - \mathbf{x}^\star)\|^2\Big] \leq \frac{2\phi_0(\mathbf{x}^\star)}{A_K} = \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{\Lambda}}^2}{A_K}.$$

When $K \geq \frac{5}{2}n\log n$, we have $A_K \geq \frac{(K - \frac{5}{2}n\log n)^2}{36n^2}$. If $\phi_0(\mathbf{x}^\star) \leq |f(\mathbf{x}^\star)|$, then for $K \geq \frac{5}{2}n\log n + \frac{6n}{\sqrt{\varepsilon}}$, we have $\mathbb{E}[f(\widetilde{\mathbf{x}}_K) - f(\mathbf{x}^\star)] \leq \varepsilon |f(\mathbf{x}^\star)|$. The total cost is $O\big(nnz(\mathbf{A})\big(\log n + \frac{1}{\sqrt{\varepsilon}}\big)\big)$.

Proof. Observe that, by the choice of step sizes, $n^2 a_{k-1}^2 \leq \frac{A_{k-2}}{4}$, $\forall k \geq 3$. Thus, telescoping the bound in Lemma 3.3 and combining with Lemma 3.4, we have

$$\begin{aligned} &\mathbb{E}[A_K \mathrm{G}_K(\mathbf{u}, \mathbf{v})] \\ \leq &\phi_0(\mathbf{u}) - \frac{A_K}{2}\mathbb{E}[\|\mathbf{v} - \mathbf{y}_K\|_2^2] - \frac{1}{2}\mathbb{E}[\|\mathbf{u} - \mathbf{x}_K\|_{\mathbf{\Lambda}}^2] - a_K\mathbb{E}[\langle \mathbf{A}(\mathbf{u} - \mathbf{x}_K), \mathbf{y}_K - \mathbf{y}_{K-1}\rangle] \\ &- \frac{A_{K-1}}{4}\mathbb{E}[\|\mathbf{y}_K - \mathbf{y}_{K-1}\|_2^2] + n^2 {a_1}^2 \mathbb{E}[\|\mathbf{y}_1 - \mathbf{y}_0\|_2^2] - \mathbb{E}[\phi_0(\mathbf{x}_1)]. \end{aligned} \tag{38}$$

We first show how to cancel out the inner product term with the negative quadratic terms. Observe that, $\forall \beta > 0$,

$$\begin{aligned}-a_K\langle \mathbf{A}(\mathbf{u}-\mathbf{x}_K), \mathbf{y}_K - \mathbf{y}_{K-1}\rangle =& -a_K \sum_{j=1}^{n} (\mathbf{y}_K - \mathbf{y}_{K-1})^\top \mathbf{A}_{:j}(u_j - [\mathbf{x}_K]_j) \\ \leq& a_K\Big(\frac{n\beta}{2}\|\mathbf{y}_K - \mathbf{y}_{K-1}\|_2^2 + \frac{1}{2\beta}\|\mathbf{u}-\mathbf{x}_K\|_{\mathbf{\Lambda}}^2\Big),\end{aligned}$$

where the last line is by Young's inequality. In particular, choosing $\beta = 2a_K$, we have $\frac{1}{2}a_K n\beta = na_K^2$, which is at most $\frac{A_{K-1}}{4}$, by the choice of step sizes in SI-NNLS+. Thus, since $\phi_0(\mathbf{x}_1) = \frac{1}{2}\|\mathbf{x}_1 - \mathbf{x}_0\|_{\mathbf{\Lambda}}^2$, Eq. (38) simplifies to

$$\begin{aligned}&\mathbb{E}[A_K \mathrm{G}_K(\mathbf{u},\mathbf{v})] \\ \leq& \phi_0(\mathbf{u}) - \frac{A_K}{2}\mathbb{E}[\|\mathbf{v}-\mathbf{y}_K\|_2^2] - \frac{1}{4}\mathbb{E}[\|\mathbf{u}-\mathbf{x}_K\|_{\mathbf{\Lambda}}^2] + n^2 {a_1}^2 \mathbb{E}[\|\mathbf{y}_1 - \mathbf{y}_0\|_2^2] - \mathbb{E}[\phi_0(\mathbf{x}_1)].\end{aligned} \tag{39}$$

Since $-\frac{1}{4}\mathbb{E}[\|\mathbf{u}-\mathbf{x}_K\|_{\mathbf{\Lambda}}^2] \leq 0$, we can ignore it. Let us now bound $n^2 {a_1}^2\|\mathbf{y}_1 - \mathbf{y}_0\|_2^2 - \phi_0(\mathbf{x}_1)$. By definition of $\mathbf{x}_1$ and ϕ_1, we have $\mathbf{x}_1 = \mathbf{x}_0 - a_1\mathbf{\Lambda}^{-1}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})$. Further, from Eq. (9) and Eq. (6), as we have $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$ and $\mathbf{y}_0 = \mathbf{A}\mathbf{x}_0$, we can simplify the terms to bound as follows.

$$\begin{aligned}&n^2 a_1^2\|\mathbf{y}_1 - \mathbf{y}_0\|_2^2 - \phi_0(\mathbf{x}_1) \\ =& {a_1}^2\Big(n^2 {a_1}^2\|\mathbf{A}\mathbf{\Lambda}^{-1}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})\|_2^2 - \frac{1}{2}\|\mathbf{\Lambda}^{-1}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})\|_{\mathbf{\Lambda}}^2\Big) \\ \leq& {a_1}^2\Big(n^2 {a_1}^2\|\mathbf{\Lambda}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{\Lambda}^{-1/2}\|_2\|\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})\|_2^2 - \frac{1}{2}\|\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})\|_2^2\Big) \\ \leq& {a_1}^2\|\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A}^\top\mathbf{y}_0 - \mathbf{1})\|_2^2\Big(n^3 {a_1}^2 - \frac{1}{2}\Big),\end{aligned}$$

where the reasoning behind the first inequality follows from the definition of spectral norm and that $\|\mathbf{A}\mathbf{\Lambda}^{-1/2}\|_2^2 = \lambda_{\max}(\mathbf{\Lambda}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{\Lambda}^{-1/2})$; the last inequality follows as the matrix $\mathbf{\Lambda}^{-1/2}\mathbf{A}^\top\mathbf{A}\mathbf{\Lambda}^{-1/2}$ has all ones on the main diagonal, and thus its trace is at most n, and since it is positive semidefinite, its spectral norm is at most its trace. As $a_1 = \frac{1}{\sqrt{2}n^{1.5}}$, we conclude that $n^2 {a_1}^2\|\mathbf{y}_1 - \mathbf{y}_0\|_2^2 - \phi_0(\mathbf{x}_1) \leq 0$. Thus, Eq. (39) simplifies to

$$\mathbb{E}[A_K \mathrm{G}_K(\mathbf{u},\mathbf{v})] \leq \mathbb{E}[\phi_0(\mathbf{u})] - \frac{A_K}{2}\mathbb{E}[\|\mathbf{v}-\mathbf{y}_K\|_2^2]. \tag{40}$$

By construction, $\mathrm{Gap}_{\mathcal{L}}^{\mathbf{u},\mathbf{v}}(\widetilde{\mathbf{x}}_K, \widetilde{\mathbf{y}}_K) \leq \mathrm{G}_K(\mathbf{u},\mathbf{v})$. Further, by Inequality (5), $f(\widetilde{\mathbf{x}}_K) - f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{A}(\widetilde{\mathbf{x}}_K - \mathbf{x}^\star)\|^2 \leq \mathrm{Gap}_{\mathcal{L}}^{\mathbf{u},\mathbf{v}}(\widetilde{\mathbf{x}}_K, \widetilde{\mathbf{y}}_K)$. Hence, we can conclude from Eq. (40) that

$$\mathbb{E}\Big[f(\widetilde{\mathbf{x}}_K) - f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{A}(\widetilde{\mathbf{x}}_K - \mathbf{x}^\star)\|^2\Big] \leq \frac{\phi_0(\mathbf{x}^\star)}{A_K} = \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{\Lambda}}^2}{2A_K}. \tag{41}$$

On the other hand, for $\mathbf{u} = \mathbf{x}^\star$ and $\mathbf{v} = \mathbf{y}^\star = \mathbf{A}\mathbf{x}^\star$, $\mathrm{Gap}_{\mathcal{L}}^{\mathbf{u},\mathbf{v}}(\widetilde{\mathbf{x}}_K, \widetilde{\mathbf{y}}_K) \geq 0$, and, recalling from Eq. (6), Eq. (8), and Eq. (9) that $\mathbf{y}_K = \mathbf{A}\widetilde{\mathbf{x}}_K$, we can also conclude from Eq. (40) that

$$\mathbb{E}\Big[\frac{1}{2}\|\mathbf{A}(\widetilde{\mathbf{x}}_K - \mathbf{x}^\star)\|_2^2\Big] \leq \frac{\phi_0(\mathbf{x}^\star)}{A_K} = \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{\Lambda}}^2}{2A_K}. \tag{42}$$

By Proposition 2.1 (b), $f(\mathbf{x}^\star) = -\frac{1}{2}\mathbf{1}^\top\mathbf{x}^\star = -\frac{1}{2}\|\mathbf{A}\mathbf{x}^\star\|_2^2$. Using this identity, one can verify that, $\forall \mathbf{x}$,

$$\begin{aligned}f(\mathbf{x}) - f(\mathbf{x}^\star) + \frac{1}{2}\|\mathbf{A}(\mathbf{x}-\mathbf{x}^\star)\|_2^2 &= \langle \mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{1}, \mathbf{x} - \mathbf{x}^\star\rangle \\ &= \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle.\end{aligned}$$

Hence, summing Eq. (41) and Eq. (42), we also have

$$\mathbb{E}\Big[\langle \nabla f(\widetilde{\mathbf{x}}_K), \widetilde{\mathbf{x}}_K - \mathbf{x}^\star\rangle + \frac{1}{2}\|\mathbf{A}(\widetilde{\mathbf{x}}_K - \mathbf{x}^\star)\|^2\Big] \leq \frac{2\phi_0(\mathbf{x}^\star)}{A_K} = \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{\Lambda}}^2}{A_K}.$$

Finally, the bound on the rate of growth of A_k is provided in Appendix 6.3. □

The reason $\mathbf{A}$ does not show up in the final bounds (thereby rendering our algorithm "scale-invariant") is because Proposition 2.1 allows bounding $\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{A}}^2$ by $|f(\mathbf{x}^\star)|$ where we crucially use the non-negativity of $\mathbf{A}$ and $\mathbf{x}$. This does not seem possible for general $\mathbf{A}$. However, an additive (as opposed to multiplicative) error bound can be obtained even with the more general $\mathbf{A}$ with only small updates to the analysis. This bound would necessarily depend on the scale of $\mathbf{A}$. The choice of the regularizer $\frac{1}{2}\|\cdot - \mathbf{x}_0\|_{\mathbf{A}}^2$ is also crucial here.

6.3 Omitted Proofs from Section 3: Growth of Scalar Sequences

In this section, we use the properties of $\{a_i\}$ and $\{a_i^k\}$ to obtain our claimed rate of growth of A_k. Note that in any iteration $k \geq 2$ of Algorithm 1, there are two possible updates to a_k, which we name as follows.

$$\text{Type I update: } a_{k+1} = \frac{n a_k}{n-1} \tag{43}$$

$$\text{Type II update: } a_{k+1} = \frac{\sqrt{A_k}}{2n} \tag{44}$$

Obtaining a handle on the growth rate of A_k requires controlling the number of updates of both types specified above. At a high level, the idea behind obtaining such a bound is that if the algorithm had only Type II updates, we would have $A_k \geq \Omega(\frac{k^2}{n^2})$; we then go on to show that we cannot have more than $\frac{5}{2} n \log n$ Type I updates since those make a_k grow too fast. We formalize this intuition in the following lemmas.

Lemma 6.4. *In Algorithm 1, we have, for $k \geq 2$, that $a_{k+1} \geq a_k$ and $A_{k+1} > A_k$.*

Proof. Notice that for all k, we have $a_k > 0$, which implies that $A_{k+1} \overset{\text{def}}{=} A_k + a_{k+1}$ satisfies $A_{k+1} > A_k$. To check the non-decreasing nature of a_k, we recall that $a_{k+1} = \min\left(\frac{n a_k}{n-1}, \frac{\sqrt{A_k}}{2n}\right)$. In the case that $\frac{\sqrt{A_k}}{2n} \geq \frac{n a_k}{n-1}$, we have $a_{k+1} = \frac{n a_k}{n-1} > a_k$, as claimed. Consider the other case with $a_{k+1} = \frac{\sqrt{A_k}}{2n}$, and suppose, for the sake of contradiction, that $a_{k+1} < a_k$. Chaining this inequality with the assumed expression for a_{k+1}, scaling appropriately, and squaring both sides gives $A_k < 4n^2 a_k^2$. Plugging this into $A_k = A_{k-1} + a_k$ and solving for a_k from this quadratic inequality (and further invoking the nonnegativity of a_k), yields $a_k > \frac{1+\sqrt{1+16n^2 A_{k-1}}}{8n^2} > \frac{\sqrt{A_{k-1}}}{2n}$. However, this contradicts $a_k = \min\left(\frac{n a_{k-1}}{n-1}, \frac{\sqrt{A_{k-1}}}{2n}\right) \leq \frac{\sqrt{A_{k-1}}}{2n}$. □

Lemma 6.5. *Consider the iterations $\{s_k\}$ in which Algorithm 1 performs a Type II update $a_{s_k+1} = \frac{\sqrt{A_{s_k}}}{2n}$. Then we have $A_{s_k} \geq \frac{k^2}{c_1 n^2}$ and $a_{s_k} \geq \frac{k-1}{2\sqrt{c_1} n^2}$ for $c_1 = 36$.*

Proof. We prove this claim by induction. First, notice that $s_k \geq k+1$ for any k. Recall our initialization $a_1 = A_1 = \frac{1}{\sqrt{2} n^{1.5}}$. By combining this with the monotonicity property stated in Lemma 6.4, we have $a_{s_1} \geq a_2 = \frac{1}{\sqrt{2} n^{2.5}} \geq 0$. By using Lemma 6.4 again, we have, in a similar fashion, that $A_{s_1} \geq A_2 = \frac{1}{\sqrt{2} n^{2.5}} + \frac{1}{\sqrt{2} n^{1.5}} \geq \frac{1}{c_1 n^2}$, which proves the base case for induction. Assume that for some $k > 1$, we have the induction hypothesis $A_{s_k} \geq \frac{k^2}{c_1 n^2}$ and $a_{s_k} \geq \frac{k-1}{2\sqrt{c_1} n^2}$. Then, combining the monotonicity of A_k from Lemma 6.4 with the fact that the algorithm performs a Type II update on a_{s_k}, we have $a_{s_{k+1}} = \frac{\sqrt{A_{s_{k+1}-1}}}{2n} \geq \frac{\sqrt{A_{s_k}}}{2n} \geq \frac{k}{2\sqrt{c_1} n^2}$. By again applying monotonicity of A_k and the induction hypothesis about a_k, we have $A_{s_{k+1}} = A_{s_{k+1}-1} + a_{s_{k+1}} \geq A_{s_k} + a_{s_{k+1}} \geq \frac{2k^2 + \sqrt{c_1} k}{2 c_1 n^2} > \frac{(k+1)^2}{c_1 n^2}$. □

Lemma 6.6. *If at some k_0^{th} iteration of Algorithm 1, we have that*

$$a_{k_0} > \frac{n-1}{2\sqrt{c_1} n^2}; \quad A_{k_0} \geq \frac{1}{c_1} \tag{45}$$

then for all $k \geq k_0$, we have that

$$a_k \geq \frac{k-1-k_0+n}{2\sqrt{c_1}n^2}; \quad A_k \geq \frac{(k-k_0+n)^2}{c_1 n^2} \tag{46}$$

for $c_1 = 36$.

Proof. We prove the claim by induction. First, the base case is true for $k = k_0$ by our assumption on a_{k_0} and A_{k_0}. Assume the induction hypothesis $a_k \geq \frac{k-1-k_0+n}{2\sqrt{c_1}n^2}$ and $A_k \geq \frac{(k-k_0+n)^2}{c_1 n^2}$ for $k \geq k_0$. We now discuss how a_k changes with the two types of updates.

If the algorithm performs a Type I update on a_k, then, by definition, $a_{k+1} = \frac{na_k}{n-1}$. Now applying the assumed lower bound on a_k, we have, when $k > k_0$, that

$$a_{k+1} = \frac{na_k}{n-1} \geq \frac{k-1-k_0+n}{2\sqrt{c_1}n(n-1)} \geq \frac{k-k_0+n}{2\sqrt{c_1}n^2}.$$

Similarly, given that $A_k \geq \frac{(k-k_0+n)^2}{c_1 n^2}$, we have,

$$A_{k+1} = A_k + a_{k+1} \geq \frac{(k-k_0+n)^2}{c_1 n^2} + \frac{k-k_0+n}{2\sqrt{c_1}n^2} \geq \frac{(k+1-k_0+n)^2}{c_1 n^2}.$$

If, on the other hand, the algorithm performs a Type II update on a_k, then we have

$$a_{k+1} = \frac{\sqrt{A_k}}{2n} \geq \frac{k-k_0+n}{2\sqrt{c_1}n^2}.$$

This completes the induction. □

As we saw in Lemma 6.6, after the k_0^{th} iteration - starting at which Inequality (45) holds - A_k grows fast. We therefore need to estimate the number of Type I updates *before* the k_0^{th} iteration.

Lemma 6.7. *There are at most $\frac{3}{2}n\log n$ Type I updates (Equation (43)) performed before the k_0^{th} iteration (the first iteration at which Inequality (45) holds).*

Proof. Suppose there are n_1 Type I updates performed by Algorithm 1 before the k_0^{th} iteration, when Inequality (45) starts to hold. Further, by Lemma 6.4, a_k is monotonically increasing (for both types of updates). Then, when considering Type I updates (Equation (43)), we have $a_{k_0} \geq \left(\frac{n}{n-1}\right)^{n_1} a_2 = \left(\frac{n}{n-1}\right)^{n_1} \cdot \frac{1}{\sqrt{2}n^{2.5}}$. In order for $a_{k_0} > \frac{n-1}{12n^2}$, we only need to have $n_1 > \log_{\frac{n}{n-1}}\left(\frac{\sqrt{n}(n-1)}{6\sqrt{2}}\right)$. In a similar fashion, combining the monotonicity of A_k from Lemma 6.4 with the Type I update rule, we have

$$A_{k_0} \geq a_2\left(1 + \frac{n}{n-1} + \left(\frac{n}{n-1}\right)^2 + \cdots + \left(\frac{n}{n-1}\right)^{n_1}\right) > \frac{(\frac{n}{n-1})^{n_1}}{\sqrt{2}n^{2.5}}.$$

So, in order to have $A_{k_0} > \frac{1}{36}$ per Inequality (45), we only need to have $n_1 \geq \log_{\frac{n}{n-1}}(\frac{n^{2.5}}{18\sqrt{2}})$. By using the approximation $1 + x \leq e^x$ and combining the above two bounds, we get as soon as $n_1 \geq \frac{3}{2}n\log n$, the inequality (45) holds. □

Proposition 6.8. *[Rate of change of A_k] When $k \geq \frac{5}{2}n\log n$, we have $A_k \geq \frac{(k-\frac{5}{2}n\log n)^2}{36n^2}$.*

Proof. Let there be t_1 Type I updates and t_2 Type II updates before the first iteration at which Inequality (45) holds, and let us call this iteration k_0. By the result of Lemma 6.7, we have $t_1 \leq \frac{3}{2}n\log n$. By the result of Lemma 6.5, we must have $A_{k_0} \geq \frac{t_2^2}{c_1 n^2}$ and $a_{k_0} \geq \frac{t_2-1}{2\sqrt{c_1}n^2}$. To meet the requirement in Inequality (45) then, we can see that $t_2 \leq n$. Therefore, $k_0 = t_1 + t_2 \leq \frac{3}{2}n\log n + n \leq \frac{5}{2}n\log n$. Having reached the k_0^{th} iteration, the result of Lemma 6.6 applies, and we have $A_k \geq \frac{(k-k_0)^2}{c_1 n^2}$. □

6.4 Omitted Proofs from Section 4: Restart Strategy

To establish local error bounds, we start with the observation that (P) is equivalent to a linear complementarity problem.

Proposition 6.9. *Problem* (P) *is equivalent to the following linear complementarity problem, denoted by* $\mathrm{LCP}(\mathbf{M}, \mathbf{q})$.

$$\mathbf{M}\mathbf{x} + \mathbf{q} \geq \mathbf{0},\ \mathbf{x} \geq \mathbf{0},\ \langle \mathbf{x}, \mathbf{M}\mathbf{x} + \mathbf{q} \rangle = 0, \tag{47}$$

where $\mathbf{\Lambda}^{-1}\mathbf{M} = \mathbf{A}^\top \mathbf{A}$ *and* $\mathbf{q} = -\mathbf{\Lambda}^{-1}\mathbf{1}$.

Proof. Observe first that, as $\mathbf{\Lambda}^{-1}$ is a diagonal matrix with positive elements on the diagonal, the stated linear complementarity problem is equivalent to

$$\nabla f(\mathbf{x}) \geq \mathbf{0},\ \mathbf{x} \geq \mathbf{0},\ \langle \nabla f(\mathbf{x}), \mathbf{x} \rangle = 0. \tag{48}$$

By Proposition 2.1, these conditions hold for any solution of (P). In the opposite direction, suppose that the conditions from Eq. (48) hold for some $\mathbf{x}$. Then applying these conditions for any $\mathbf{u} \geq \mathbf{0}$ gives

$$\langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle = \langle \nabla f(\mathbf{x}), \mathbf{u} \rangle \geq 0.$$

But $\langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle \geq 0$ is the first-order optimality condition for (P), and so $\mathbf{x}$ solves (P). □

For $r(\mathbf{x}) = \|\mathbf{R}(\mathbf{x})\|_{\mathbf{\Lambda}}$, a quantity termed *natural residual* [31], local error bound is obtained as a corollary of the following theorem.

Theorem 6.10 ([31], Theorem 2.1)**.** *Let* $\mathbf{M} \in \mathbb{R}^{n \times n}$ *be such that* $\mathrm{LCP}(\mathbf{M}, \mathbf{0})$ *has* $\mathbf{0}$ *as its unique solution. Then there exists* $\mu > 0$ *such that for each* $\mathbf{x} \in \mathbb{R}^n$*, we have* $r(\mathbf{x}) \geq \mu \|\mathbf{x} - \mathbf{x}^\star\|$*, where* $\mathbf{x}^\star$ *is a solution to* $\mathrm{LCP}(\mathbf{M}, \mathbf{q})$ *that is closest to* $\mathbf{x}$ *under the norm* $\|\cdot\|$.

Theorem 6.10 applies to our problem due to the nonnegativity (and nondegeneracy) of $\mathbf{A}$ and choosing $\|\cdot\| = \|\cdot\|_{\mathbf{\Lambda}}$. By arguing that Theorem 3.5 provides an upper bound on $r(\widetilde{\mathbf{x}}_K)$, in expectation, we then obtain our final result below.

Proposition 6.11. *For any* $\mathbf{x} \in \mathbb{R}^n_+$, $r(\mathbf{x}) \leq \sqrt{2n(f(\mathbf{x}) - f(\mathbf{x}^\star))}$, *where* $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^n_+} f(\mathbf{u})$.

Proof. Given $\mathbf{x} \in \mathbb{R}^n_+$, consider $\hat{\mathbf{x}}$ defined as $\hat{x}_{j^\star} = x_{j^\star} - \mathbf{R}_{j^\star}(\mathbf{x})$, where $j^\star = \operatorname{argmax}_{1 \leq j \leq n} |\mathbf{R}_j(\mathbf{x})| \cdot \|\mathbf{A}_{:j}\|_2$, and $\hat{x}_j = x_j$ for $j \neq j^\star$. Then observing that

$$\begin{aligned} f(\hat{\mathbf{x}}) - f(\mathbf{x}) &= \nabla_{j^\star} f(\mathbf{x})([\hat{\mathbf{x}}]_{j^\star} - [\mathbf{x}]_{j^\star}) + \frac{\|\mathbf{A}_{:j^\star}\|_2^2}{2} |[\hat{\mathbf{x}}]_{j^\star} - [\mathbf{x}]_{j^\star}|^2 \\ &\leq -\frac{1}{2} |\mathbf{R}_{j^\star}(\mathbf{x})|^2 \|\mathbf{A}_{:j^\star}\|_2^2 \leq -\frac{1}{2n} \|\mathbf{R}(\mathbf{x})\|_{\mathbf{\Lambda}}^2, \end{aligned}$$

and combining with $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}^\star)$, $r(\mathbf{x}) = \|\mathbf{R}(\mathbf{x})\|_{\mathbf{\Lambda}}$, the claimed bound follows after a simple rearrangement. □

Theorem 4.1. *Given an error parameter* $\varepsilon > 0$ *and* $\mathbf{x}_0 = \mathbf{0}$*, consider the following algorithm* $\mathcal{A}$:

> $\mathcal{A}$: ***SI-NNLS+ with Restarts***
> *Initialize:* $k = 1$.
> *Initialize Lazy SI-NNLS+ at* $\mathbf{x}_{k-1}$.
> *Run Lazy SI-NNLS+ until the output* $\widetilde{\mathbf{x}}_K^k$ *satisfies* $r(\widetilde{\mathbf{x}}_K^k) \leq \frac{1}{2} r(\mathbf{x}_{k-1})$.
> *Restart Lazy SI-NNLS+ initializing at* $\mathbf{x}_k = \widetilde{\mathbf{x}}_K^k$.
> *Increment* k.
> *Repeat until* $r(\widetilde{\mathbf{x}}_K^k) \leq \varepsilon$.

Then, the expected number of arithmetic operations of $\mathcal{A}$ *is* $O\big(\mathrm{nnz}(\mathbf{A})\big(\log n + \frac{\sqrt{n}}{\mu}\big) \log\big(\frac{r(\mathbf{x}_0)}{\varepsilon}\big)\big)$. *As a consequence, given* $\bar{\varepsilon} > 0$*, the total expected number of arithmetic operations until a point with* $f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \bar{\varepsilon} |f(\mathbf{x}^\star)|$ *can be constructed by* $\mathcal{A}$ *is* $O\Big(\mathrm{nnz}(\mathbf{A})\Big(\log n + \frac{\sqrt{n}}{\mu}\Big) \log\Big(\frac{n}{\mu\bar{\varepsilon}}\Big)\Big)$.

Proof. Because each restart halves the natural residual $r(\mathbf{x})$, it is immediate that the total number of restarts until $r(\widetilde{\mathbf{x}}_K^k) \leq \varepsilon$ is bounded by $\log(\frac{r(\mathbf{x}_0)}{\varepsilon})$. Thus, to prove the first (and main) part of the theorem, we only need to bound the number of iterations (and the overall number of arithmetic operations) of (Lazy) SI-NNLS+ in expectation. Hence, in the following, we only consider one run of SI-NNLS+ until the natural residual is halved. To keep the notation simple, we let $\mathbf{x}_0$ denote the initial point of SI-NNLS+ and $\widetilde{\mathbf{x}}_k$ denote the output of SI-NNLS+ at iteration k. If $r(\mathbf{x}_0) = 0$, $\mathcal{A}$ halts immediately and the bound on the number of iterations holds trivially, so assume $r(\mathbf{x}_0) > 0$. Using Theorem 3.5, we have that $\forall k \geq 2$,

$$\mathbb{E}[A_k r^2(\widetilde{\mathbf{x}}_k)] \leq n\|\mathbf{x}_0 - \mathbf{x}^\star\|_{\mathbf{\Lambda}}^2 \leq \frac{n}{\mu^2} r^2(\mathbf{x}_0). \tag{49}$$

As $r^2(\cdot)$ is nonnegative, we can use Markov's inequality to bound the total number of iterations K until $r(\widetilde{\mathbf{x}}_K) \leq \frac{r(\mathbf{x}_0)}{2}$. In particular, using Eq. (49), we get by Markov's inequality that $\Pr[K > k] \leq \Pr[r^2(\widetilde{\mathbf{x}}_k) \geq \frac{r^2(\mathbf{x}_0)}{4}] \leq \frac{4n}{\mu^2 A_k}$. As K is nonnegative, we can estimate its expectation using

$$\begin{aligned}
\mathbb{E}[K] &= \sum_{i=0}^{\infty} \Pr[K > i] \leq \sum_{i=0}^{\infty} \min\left\{1, \frac{4n}{\mu^2 A_i}\right\} \\
&\leq \sum_{i=0}^{\lceil 12n\sqrt{n}/\mu + \frac{5}{2}n\log n\rceil} 1 + \sum_{\lceil 12n\sqrt{n}/\mu + \frac{5}{2}n\log n\rceil + 1}^{\infty} \frac{4n}{\mu^2 A_i} \\
&\leq 24n\sqrt{n}/\mu + \frac{5}{2}n\log n + 2,
\end{aligned}$$

where in the last inequality we use the rate of A_k from Proposition 6.8.

In the lazy implementation of SI-NNLS+, as argued in Appendix 7, the expected cost of an iteration is $\frac{\mathrm{nnz}(\mathbf{A})}{n}$, which leads to the claimed bound on the number of arithmetic operations until $r(\mathbf{x}) \leq \varepsilon$.

By using that $r(\mathbf{x}_0) \leq \sqrt{2n(f(\mathbf{x}_0) - f(\mathbf{x}^\star))} = \sqrt{2n|f(\mathbf{x}^\star)|}$, $f\left(\widetilde{\mathbf{x}}^{K_1} - \mathbf{R}(\widetilde{\mathbf{x}}^{K_1})\right) - f(\mathbf{x}^\star) \leq \left((n-1) + \frac{n+1}{\mu}\right) r^2(\widetilde{\mathbf{x}}^{K_1})$ (argued below), the bound on the number of iterations until $f\left(\widetilde{\mathbf{x}}^{K_1} - \mathbf{R}(\widetilde{\mathbf{x}}^{K_1})\right) - f(\mathbf{x}^\star) \leq \bar{\varepsilon}|f(\mathbf{x}^\star)|$ have

$$\begin{aligned}
f\left(\widetilde{\mathbf{x}}^{K_1} - \mathbf{R}(\widetilde{\mathbf{x}}^{K_1})\right) - f(\mathbf{x}^\star) &\leq \left((n-1) + \frac{n+1}{\mu}\right) r^2(\widetilde{\mathbf{x}}^{K_1}) \\
&\leq \left((n-1) + \frac{n+1}{\mu}\right) \frac{1}{2^{2K_1}} r^2(\mathbf{x}_0) \\
&\leq \left((n-1) + \frac{n+1}{\mu}\right) \frac{1}{2^{2K_1}} 2n|f(\mathbf{x}^\star)|
\end{aligned}$$

and by setting $K_1 = \frac{1}{2}\log_2 \frac{2n\left((n-1)+\frac{n+1}{\mu}\right)}{\bar{\varepsilon}}$, we have this bound.

Finally, it remains to argue that $f\left(\widetilde{\mathbf{x}}^{K_1} - \mathbf{R}(\widetilde{\mathbf{x}}^{K_1})\right) - f(\mathbf{x}^\star) \leq \left((n-1) + \frac{n+1}{\mu}\right) r^2(\widetilde{\mathbf{x}}^{K_1})$. Observe that the definition of $\mathbf{R}(\mathbf{x})$ is equivalent to $\mathbf{x} - \bar{\mathbf{x}}$, where

$$\bar{\mathbf{x}} = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^n_+} \left\{ \langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|_{\mathbf{\Lambda}}^2 \right\}.$$

By the first-order optimality of $\bar{\mathbf{x}}$ based on the equivalent definition of $\mathbf{R}(\mathbf{x})$ above, we have $\langle \nabla f(\mathbf{x}) + \mathbf{\Lambda}(\bar{\mathbf{x}} - \mathbf{x}), \mathbf{x}^\star - \bar{\mathbf{x}}\rangle \geq 0$. Rearranging, and using the definition of convexity of f, we have

$$\begin{aligned}
f(\bar{\mathbf{x}}) - f(\mathbf{x}^\star) &\leq \langle \nabla f(\bar{\mathbf{x}}), \bar{\mathbf{x}} - \mathbf{x}^\star\rangle \\
&\leq \langle \nabla f(\mathbf{x}) - \nabla f(\bar{\mathbf{x}}) + \mathbf{\Lambda}(\bar{\mathbf{x}} - \mathbf{x}), \mathbf{x}^\star - \bar{\mathbf{x}}\rangle \\
&= \langle (\mathbf{A}^\top\mathbf{A} - \mathbf{\Lambda})(\mathbf{x} - \bar{\mathbf{x}}), \mathbf{x}^\star - \bar{\mathbf{x}}\rangle \\
&= \langle (\mathbf{A}^\top\mathbf{A} - \mathbf{\Lambda})\mathbf{R}(\mathbf{x}), \mathbf{R}(\mathbf{x})\rangle + \langle (\mathbf{A}^\top\mathbf{A} - \mathbf{\Lambda})\mathbf{R}(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle \\
&= \langle (\mathbf{A}^\top\mathbf{A} - \mathbf{\Lambda})\mathbf{R}(\mathbf{x}), \mathbf{R}(\mathbf{x})\rangle + \langle \mathbf{A}^\top\mathbf{A}\mathbf{R}(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle - \langle \mathbf{\Lambda}\mathbf{R}(\mathbf{x}), \mathbf{x}^\star - \mathbf{x}\rangle \\
&\leq (n-1)\|\mathbf{R}(\mathbf{x})\|_{\mathbf{\Lambda}}^2 + (n+1)\|\mathbf{R}(\mathbf{x})\|_{\mathbf{\Lambda}}\|\mathbf{x} - \mathbf{x}^\star\|_{\mathbf{\Lambda}} \\
&\leq \left((n-1) + \frac{n+1}{\mu}\right) r^2(\mathbf{x}),
\end{aligned}$$

where in the last inequality we have used the error bound from Theorem 6.10. □

7 Implementation Version of SI-NNLS+

Since Algorithm 1 explicitly updates $\widetilde{\mathbf{x}}_k$ and $\widetilde{\mathbf{y}}_k$ (of lengths n and m respectively), the per iteration cost is $O(m+n)$, which is unnecessarily high when the matrix $\mathbf{A}$ is sparse. In this section, we show that by using a *lazy* update strategy, we can efficiently implement Algorithm 1 with overall complexity independent of the ambient dimension. To attain this result, we maintain implicit representations for $\widetilde{\mathbf{x}}_k$, $\mathbf{y}_k$, and $\bar{\mathbf{y}}_k$ by introducing two auxiliary variables that are amenable to efficient updates.

Efficiently Updating the Primal Variable. In Lemma 7.1, we show that we can work with an implicit representation of $\widetilde{\mathbf{x}}_k$ by introducing $\mathbf{r}_k$.

Lemma 7.1. *For $\{\widetilde{\mathbf{x}}_k\}$ defined in Eq. (8) (and simplified in Algorithm 1), we have, for $k \geq 1$,*

$$\widetilde{\mathbf{x}}_k = \mathbf{x}_k + \frac{1}{A_k}\mathbf{r}_k, \tag{50}$$

where $\mathbf{x}_k$ evolves as per Algorithm 1, $\mathbf{r}_1 = \mathbf{0}$ and, when $k \geq 1$, $\mathbf{r}_k = \mathbf{r}_{k-1} + ((n-1)a_k - A_{k-1})(\mathbf{x}_k - \mathbf{x}_{k-1})$.

Proof. We prove the lemma by induction. Using the facts that $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{x}_1 = \widetilde{\mathbf{x}}_1$, $\mathbf{s}_1 = \mathbf{0}$, $a_1 = A_1$, and $A_0 = 0$, we have

$$\begin{aligned}\widetilde{\mathbf{x}}_1 &= \frac{1}{A_1}\left(A_0\widetilde{\mathbf{x}}_0 + a_1\big(n\mathbf{x}_1 - (n-1)\mathbf{x}_0\big)\right)\\ &= \frac{1}{A_1}\left(a_1\mathbf{x}_1 + (n-1)a_1(\mathbf{x}_1 - \mathbf{x}_0)\right)\\ &= \mathbf{x}_1 + ((n-1)a_1 - A_0)(\mathbf{x}_1 - \mathbf{x}_0).\end{aligned} \tag{51}$$

Assume for certain $k \geq 2$, that Eq. (50) holds for $k-1$. Then, using the recursion of $\tilde{\mathbf{x}}_k$ in Algorithm 1, we have that for $k \geq 3$,

$$\begin{aligned}A_k\widetilde{\mathbf{x}}_k &= A_{k-1}\widetilde{\mathbf{x}}_{k-1} + a_k\mathbf{x}_k + (n-1)a_k(\mathbf{x}_k - \mathbf{x}_{k-1})\\ &= A_{k-1}\mathbf{x}_{k-1} + \mathbf{r}_{k-1} + a_k\mathbf{x}_k + (n-1)a_k(\mathbf{x}_k - \mathbf{x}_{k-1})\\ &= A_{k-1}(\mathbf{x}_{k-1} - \mathbf{x}_k + \mathbf{x}_k) + \mathbf{r}_{k-1} + a_k\mathbf{x}_k + (n-1)a_k(\mathbf{x}_k - \mathbf{x}_{k-1})\\ &= A_{k-1}(\mathbf{x}_{k-1} - \mathbf{x}_k) + A_{k-1}\mathbf{x}_k + \mathbf{r}_{k-1} + a_k\mathbf{x}_k + (n-1)a_k(\mathbf{x}_k - \mathbf{x}_{k-1})\\ &= A_k\mathbf{x}_k + \mathbf{r}_{k-1} + ((n-1)a_k - A_{k-1})(\mathbf{x}_k - \mathbf{x}_{k-1})\\ &= A_k\mathbf{x}_k + \mathbf{r}_k,\end{aligned}$$

as required. □

The expression for $\mathbf{r}_k$ in Lemma 7.1 shows that it can be updated at cost $O(1)$ as $\mathbf{x}_k$ differs from $\mathbf{x}_{k-1}$ only at one coordinate. Therefore, by Eq. (50) we need not compute $\tilde{\mathbf{x}}_k$ in all iterations and can instead maintain $\mathbf{r}_k$. Along the same lines, we give an efficient implementation strategy for $\mathbf{y}_k$ and $\bar{\mathbf{y}}_k$ in the following discussion.

Efficiently Updating the Dual Variable. We now show how to update the dual variable efficiently.

Lemma 7.2. *Consider $\{\mathbf{y}_k\}$ and $\{\mathbf{x}_k\}$ evolving as per Algorithm 1. Then, for $k = 1$, we have $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1$; for $k \geq 2$, we have*

$$\mathbf{y}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{A}\mathbf{x}_k + \frac{(n-1)a_k}{A_k}\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}), \tag{52}$$

$$\tag{53}$$

Proof. The proof is directly from the definition of $\mathbf{y}_k$ in Algorithm 1. □

Lemma 7.3. *Consider $\{\mathbf{y}_k\}$ and $\{\mathbf{x}_k\}$ evolving as per Algorithm 1. Then for all $k \geq 1$, we have*

$$\mathbf{y}_k = \mathbf{A}\mathbf{x}_k + \frac{1}{A_k}\mathbf{s}_k, \tag{54}$$

where $\mathbf{s}_1 = 0$ and $\mathbf{s}_k = \mathbf{s}_{k-1} + ((n-1)a_k - A_{k-1})\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1})$ when $k \geq 2$.

Proof. We prove the lemma by induction. For the base case of $k = 1$, we have, by the choice of $\mathbf{s}_1 = \mathbf{0}$, that $\mathbf{y}_1 = \mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x}_1 + \frac{1}{A_1}\mathbf{s}_1$. Then for some $k \geq 2$, assume Eq. (54) holds for $k-1$, then we have,

$$\begin{aligned}
A_k\mathbf{y}_k &= A_{k-1}\mathbf{y}_{k-1} + a_k\mathbf{A}\mathbf{x}_k + (n-1)a_k\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= A_{k-1}\mathbf{A}\mathbf{x}_{k-1} + \mathbf{s}_{k-1} + a_k\mathbf{A}\mathbf{x}_k + (n-1)a_k\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= A_{k-1}\mathbf{A}(\mathbf{x}_{k-1} - \mathbf{x}_k + \mathbf{x}_k) + \mathbf{s}_{k-1} + a_k\mathbf{A}\mathbf{x}_k + (n-1)a_k\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= A_k\mathbf{A}\mathbf{x}_k + \mathbf{s}_{k-1} + ((n-1)a_k - A_{k-1})\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= A_k\mathbf{A}\mathbf{x}_k + \mathbf{s}_k,
\end{aligned} \tag{55}$$

where the first step is by Lemma 7.2, second step is by the induction hypothesis, third step is by adding and subtracting $A_{k-1}\mathbf{A}\mathbf{x}_k$, fourth step is by rearranging terms appropriately, and the final step uses the recursive definition of $\mathbf{s}_k$ stated in the lemma. Dividing throughout by A_k then finishes the proof. □

Algorithm 2 SI-NNLS+ (Implementation)

Input: Matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ with $n \geq 4$, accuracy ε
Output: Vector $\widetilde{\mathbf{x}}_K \in \mathbb{R}_+^n$ such that $f(\widetilde{\mathbf{x}}_K) \leq (1-\varepsilon)f(\mathbf{x}^\star)$.
Initialize: $a_1 = \frac{1}{n-1}$, $a_2 = \frac{n}{n-1}$, $A_1 = a_1$, $\phi_0(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_{\mathbf{\Lambda}}^2$, $\bar{\mathbf{y}}_0 = \mathbf{y}_0 = \mathbf{A}\mathbf{x}_0$, $\mathbf{p}_0 = \mathbf{0}, \mathbf{q}_0 = \mathbf{A}\mathbf{x}_0, \mathbf{t}_0 = \mathbf{0}, \mathbf{s}_1 = \mathbf{0}, \mathbf{r}_1 = \mathbf{0}$.
for $k = 1, 2, \ldots, K$ **do**
  Sample j_k uniformly at random from $\{1, 2, \ldots, n\}$
  if $k = 1$ **then**
    $\bar{\mathbf{y}}_0 = \mathbf{q}_0$
  else if $k = 2$ **then**
    $\bar{\mathbf{y}}_1 = \mathbf{q}_1 + \frac{a_1}{a_2}\mathbf{t}_1$
  else if $k \geq 3$ **then**
    $\bar{\mathbf{y}}_{k-1} = \mathbf{q}_{k-1} + \frac{1}{A_{k-1}}\left(1 - \frac{a_{k-1}^2}{a_k A_{k-2}}\right)\mathbf{s}_{k-1} + \frac{(n-1)a_{k-1}^2}{a_k A_{k-2}}\mathbf{t}_{k-1}$
  end if
  $p_{k,i} = \begin{cases} p_{k-1,i}, & i \neq j_k \\ p_{k-1,i} + na_k\left(\mathbf{A}_{:i}^T\bar{\mathbf{y}}_{k-1} - 1\right), & i = j_k. \end{cases}$
  $x_{k,i} = \begin{cases} x_{k-1,i}, & i \neq j_k \\ \max\left\{0, \min\left\{x_{0,i} - \frac{1}{\|\mathbf{A}_{:i}\|^2}\cdot p_{k,i}, \frac{1}{\|\mathbf{A}_{:i}\|^2}\right\}\right\}, & i = j_k \end{cases}$
  $\mathbf{t}_k = \mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1})$
  if $k \geq 2$ **then**
    $\mathbf{r}_k = \mathbf{r}_{k-1} + ((n-1)a_k - A_{k-1})(\mathbf{x}_k - \mathbf{x}_{k-1})$
    $\mathbf{s}_k = \mathbf{s}_{k-1} + ((n-1)a_k - A_{k-1})\mathbf{t}_k$
  end if
  $\mathbf{q}_k = \mathbf{q}_{k-1} + \mathbf{t}_k$
  $A_{k+1} = A_k + a_{k+1}$
  $a_{k+2} = \min\{\frac{na_{k+1}}{n-1}, \frac{\sqrt{A_{k+1}}}{2n}\}$
end for
return $\mathbf{x}_K + \frac{1}{A_K}\mathbf{r}_K$

Lemma 7.4. *Consider $\{\mathbf{x}_k\}$, $\{\mathbf{y}_k\}$, and $\{\bar{\mathbf{y}}_k\}$ evolving as per Algorithm 1. Then we have that*

$$\bar{\mathbf{y}}_1 = \mathbf{A}\mathbf{x}_1 + \frac{a_1}{a_2}\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_0). \tag{56}$$

and

$$\bar{\mathbf{y}}_k = \mathbf{A}\mathbf{x}_k + \frac{1}{A_k}\Big(1 - \frac{a_k^2}{a_{k+1}A_{k-1}}\Big)\mathbf{s}_k + \frac{(n-1)a_k^2}{a_{k+1}A_{k-1}}\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}). \tag{57}$$

Proof. From the definition of $\bar{\mathbf{y}}_k$, the initializations for $\mathbf{x}_0, \mathbf{y}_0$, and $\overline{\mathbf{y}}_0$, and Lemma 7.2, we have

$$\bar{\mathbf{y}}_1 = \mathbf{y}_1 + \frac{a_1}{a_2}(\mathbf{y}_1 - \mathbf{y}_0) = \mathbf{A}\mathbf{x}_1 + \frac{a_1}{a_2}\mathbf{A}(\mathbf{x}_1 - \mathbf{x}_0).$$

For $k \geq 2$, by Lemma 7.2, we have

$$A_k\mathbf{y}_k - A_{k-1}\mathbf{y}_{k-1} = a_k\mathbf{A}\mathbf{x}_k + (n-1)a_k\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

As a result,

$$A_{k-1}(\mathbf{y}_k - \mathbf{y}_{k-1}) = a_k\mathbf{A}\mathbf{x}_k + (n-1)a_k\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) - a_k\mathbf{y}_k. \tag{58}$$

So for $k \geq 2$, it follows that

$$\begin{aligned}
\bar{\mathbf{y}}_k &= \mathbf{y}_k + \frac{a_k}{a_{k+1}}(\mathbf{y}_k - \mathbf{y}_{k-1}) \\
&= \mathbf{y}_k + \frac{a_k}{a_{k+1}}\Big(\frac{a_k}{A_{k-1}}\mathbf{A}\mathbf{x}_k + \frac{(n-1)a_k}{A_{k-1}}\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) - \frac{a_k}{A_{k-1}}\mathbf{y}_k\Big) \\
&= \Big(1 - \frac{a_k^2}{a_{k+1}A_{k-1}}\Big)\mathbf{y}_k + \frac{a_k^2}{a_{k+1}A_{k-1}}\mathbf{A}\mathbf{x}_k + \frac{(n-1)a_k^2}{a_{k+1}A_{k-1}}\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}) \\
&= \mathbf{A}\mathbf{x}_k + \frac{1}{A_k}\Big(1 - \frac{a_k^2}{a_{k+1}A_{k-1}}\Big)\mathbf{s}_k + \frac{(n-1)a_k^2}{a_{k+1}A_{k-1}}\mathbf{A}(\mathbf{x}_k - \mathbf{x}_{k-1}),
\end{aligned}$$

where the first step is by the definition of $\overline{\mathbf{y}}_k$ in Algorithm 1, the second step is by Eq. (58), the third step is by rearranging, and the final step is by Lemma 7.3. □

Based on the above lemmas, we give our efficient lazy implementation version of Algorithm 1 in Algorithm 2. In Algorithm 2, we also introduce other auxiliary variables $\mathbf{p}_k, \mathbf{q}_k$ and $\mathbf{t}_k$. Based on Lemmas 7.1-7.4, it is easy to verify the equivalence between Algorithms 1 and 2. With this implementation, by updating only the dual coordinates corresponding to the nonzero coordinates of the selected column of $\mathbf{A}$, the per-iteration cost is proportional to the number of nonzero elements of the selected row in the iteration. As a result, the overall complexity result will depend only on the number of nonzero elements of $\mathbf{A}$.