A Proofs

We now give the details for the proof of our main results, i.e., Theorems 1 and 2. Below, we outline the steps for the proof of FLAG's Theorem 1. The proof of Theorem 2 for FLARE follows the same line of reasoning. Also, we note that, in what follows, lemmas/corollaries required for the proof of Theorem 2, are given immediately after those of FLAG.

1. FLAG is essentially a combination of mirror descent and proximal gradient descent steps (Lemmas 1 and 4).
2. L_k in Algorithm 1 plays the role of an "effective gradient Lipschitz constant" in each iteration. The convergence rate of FLAG ultimately depends on $\sum_{k=1}^{T} L_k = L\sum_{k=1}^{T} \mathbf{g}_k^T S_k^{-1}\mathbf{g}_k$. (Lemma 8 and Corollary 3)
3. By picking S_k adaptively like in AdaGrad, we achieve a non-trivial upper bound for $\sum_{k=1}^{T} L_k$. (Lemma 5)
4. FLAG relies on picking an $\mathbf{x}_k$ at each iteration that satisfies an inequality involving L_k (Corollary 1). However, because L_k is not known prior to picking $\mathbf{x}_k$, we must choose an $\mathbf{x}_k$ to roughly satisfy the inequality for all possible values of L_k. We do this by picking $\mathbf{x}_k$ using binary search. (Lemmas 2 and 3 and Corollary 1)
5. Finally, we need to pick the right stepsize for each iteration. Our scheme is very similar to the one used in [1], but generalized to handle a different L_k each iteration. (Lemmas 6 and 8 as well as Corollary 3).
6. Theorem 3 combines items 1, 2 and 4, above. Finally, to prove Theorem 1, we combine Theorem 3 with items 3 and 5 above.

A.1 Proof of Theorem 1 and Theorem 2

First, we obtain the following key result (similar to [4, Lemma 2.3]) regarding the vector $\mathbf{p} = -L(\mathbf{prox}(\mathbf{x}) - \mathbf{x})$, as in Step 3 of FLAG, which is known as the *Gradient Mapping* of F on $\mathcal{C}$.

Lemma 1 (Gradient Mapping)
For any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, we have

$$F(\boldsymbol{prox}(\mathbf{x})) \leq F(\mathbf{y}) + \langle L(\boldsymbol{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle - \frac{L}{2}\|\mathbf{x} - \boldsymbol{prox}(\mathbf{x})\|_2^2,$$

where $\boldsymbol{prox}(\mathbf{x})$ is defined as in (3). In particular, $F(\boldsymbol{prox}(\mathbf{x})) \leq F(\mathbf{x}) - \frac{L}{2}\|\mathbf{x} - \boldsymbol{prox}(\mathbf{x})\|_2^2$.

Proof of Lemma 1 This result is the same as Lemma 2.3 in [4]. We bring its proof here for completeness.

For any $\mathbf{y} \in \mathcal{C}$, any sub-gradient, $\mathbf{v}$, of h at $\mathbf{prox}(\mathbf{x})$, i.e., $\mathbf{v} \in \partial h(\mathbf{prox}(\mathbf{x}))$, and by optimality of $\mathbf{prox}(\mathbf{x})$ in (3), we have

$$\begin{aligned}0 &\leq \langle \nabla f(\mathbf{x}) + \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{prox}(\mathbf{x})\rangle \\ &= \langle \nabla f(\mathbf{x}) + \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \langle \nabla f(\mathbf{x}) \\ &\quad + \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{x} - \mathbf{prox}(\mathbf{x})\rangle,\end{aligned}$$

and so

$$\begin{aligned}&\langle \nabla f(\mathbf{x}), \mathbf{prox}(\mathbf{x}) - \mathbf{x}\rangle \\ &\leq \langle \nabla f(\mathbf{x}) + \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \\ &\quad + \langle \mathbf{v}, \mathbf{x} - \mathbf{prox}(\mathbf{x})\rangle - L\|\mathbf{x} - \mathbf{prox}(\mathbf{x})\|_2^2,\end{aligned}$$

Now from L-Lipschitz continuity of ∇f as well as convexity of f and h, we get

$$\begin{aligned}&F(\mathbf{prox}(\mathbf{x})) \\ &= f(\mathbf{prox}(\mathbf{x})) + h(\mathbf{prox}(\mathbf{x})) \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{prox}(\mathbf{x}) - \mathbf{x}\rangle \\ &\quad + \frac{L}{2}\|\mathbf{prox}(\mathbf{x}) - \mathbf{x}\|_2^2 + h(\mathbf{prox}(\mathbf{x})) \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}) + \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \\ &\quad + \langle \mathbf{v}, \mathbf{x} - \mathbf{prox}(\mathbf{x})\rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{prox}(\mathbf{x})\|_2^2 \\ &\quad + h(\mathbf{prox}(\mathbf{x})) \\ &\leq f(\mathbf{y}) + \langle \mathbf{v} + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \\ &\quad + \langle \mathbf{v}, \mathbf{x} - \mathbf{prox}(\mathbf{x})\rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{prox}(\mathbf{x})\|_2^2 \\ &\quad + h(\mathbf{prox}(\mathbf{x})) \\ &= f(\mathbf{y}) + \langle L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \\ &\quad + \langle \mathbf{v}, \mathbf{y} - \mathbf{prox}(\mathbf{x})\rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{prox}(\mathbf{x})\|_2^2 \\ &\quad + h(\mathbf{prox}(\mathbf{x})) \\ &\leq F(\mathbf{y}) + \langle L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle \\ &\quad - \frac{L}{2}\|\mathbf{x} - \mathbf{prox}(\mathbf{x})\|_2^2.\end{aligned}$$

∎

The following lemma establishes the Lipschitz continuity of the **prox** operator.

Lemma 2 (Prox Operator Continuity)
$\boldsymbol{prox} : \mathbb{R}^d \to \mathbb{R}^d$ is a 2-Lipschitz continuous, that is, for any $\mathbf{x}, \mathbf{y} \in \mathcal{C}$, we have

$$\|\boldsymbol{prox}(\mathbf{x}) - \boldsymbol{prox}(\mathbf{y})\|_2 \leq 2\|\mathbf{x} - \mathbf{y}\|_2.$$

Proof of Lemma 2 By Definition (3), for any $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{z}' \in \mathcal{C}$, $\mathbf{v} \in \partial h(\mathbf{prox}(\mathbf{x}))$, and $\mathbf{w} \in \partial h(\mathbf{prox}(\mathbf{y}))$, we have

$$\begin{aligned}
&\langle \mathbf{v}, \mathbf{z} - \mathbf{prox}(\mathbf{x}) \rangle \\
&\quad \geq -\langle \nabla f(\mathbf{x}) + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{z} - \mathbf{prox}(\mathbf{x}) \rangle, \\
&\langle \mathbf{w}, \mathbf{z}' - \mathbf{prox}(\mathbf{y}) \rangle \\
&\quad \geq -\langle \nabla f(\mathbf{y}) + L(\mathbf{prox}(\mathbf{y}) - \mathbf{y}), \mathbf{z}' - \mathbf{prox}(\mathbf{y}) \rangle.
\end{aligned}$$

In particular, for $\mathbf{z} = \mathbf{prox}(\mathbf{y})$ and $\mathbf{z}' = \mathbf{prox}(\mathbf{z})$, we get

$$\begin{aligned}
&\langle \mathbf{v}, \mathbf{prox}(\mathbf{y}) - \mathbf{prox}(\mathbf{x}) \rangle \\
&\quad \geq -\langle \nabla f(\mathbf{x}) + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{prox}(\mathbf{y}) - \mathbf{prox}(\mathbf{x}) \rangle, \\
&\langle \mathbf{w}, \mathbf{prox}(\mathbf{y}) - \mathbf{prox}(\mathbf{x}) \rangle \\
&\quad \leq \langle \nabla f(\mathbf{y}) + L(\mathbf{prox}(\mathbf{y}) - \mathbf{y}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle.
\end{aligned}$$

By monotonicity of sub-gradient, we get

$$\langle \mathbf{v}, \mathbf{prox}(\mathbf{y}) - \mathbf{prox}(\mathbf{x}) \rangle \leq \langle \mathbf{w}, \mathbf{prox}(\mathbf{y}) - \mathbf{prox}(\mathbf{x}) \rangle.$$

So

$$\begin{aligned}
&\langle \nabla f(\mathbf{x}) + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle \\
&\leq \langle \nabla f(\mathbf{y}) + L(\mathbf{prox}(\mathbf{y}) - \mathbf{y}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle,
\end{aligned}$$

and as a result

$$\begin{aligned}
&\langle \nabla f(\mathbf{x}) + L(\mathbf{prox}(\mathbf{x}) - \mathbf{x}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle \\
&= \langle \nabla f(\mathbf{x}) + L\left(\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) + \mathbf{prox}(\mathbf{y}) - \mathbf{x}\right) \\
&\quad , \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle \\
&= L\|\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y})\|_2^2 \\
&\quad + \langle \nabla f(\mathbf{x}) + L(\mathbf{prox}(\mathbf{y}) - \mathbf{x}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle \\
&\leq \langle \nabla f(\mathbf{y}) + L(\mathbf{prox}(\mathbf{y}) - \mathbf{y}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle,
\end{aligned}$$

which gives

$$\begin{aligned}
&L\|\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y})\|_2^2 \\
&\leq \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) + L(\mathbf{x} - \mathbf{y}), \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y}) \rangle \\
&\leq (\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \\
&\qquad + L\|\mathbf{x} - \mathbf{y}\|_2) \, \|\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y})\|_2 \\
&\leq 2L\|\mathbf{x} - \mathbf{y}\|_2 \|\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{y})\|_2,
\end{aligned}$$

and the result follows. ∎

Using **prox** operator continuity Lemma 2, we can conclude that given any $\mathbf{y}, \mathbf{z} \in \mathcal{C}$, if $\langle \mathbf{prox}(\mathbf{y}) - \mathbf{y}, \mathbf{y} - \mathbf{z} \rangle < 0$ and $\langle \mathbf{prox}(\mathbf{z}) - \mathbf{z}, \mathbf{y} - \mathbf{z} \rangle > 0$, then there must be a $t^* \in (0, 1)$ for which $\mathbf{w} = t^*\mathbf{y} + (1 - t^*)\mathbf{z}$ gives $\langle \mathbf{prox}(\mathbf{w}) - \mathbf{w}, \mathbf{y} - \mathbf{z} \rangle = 0$. Algorithm 2 finds an approximation to $\mathbf{w}$ in $\mathcal{O}(\log L/\epsilon)$ iterations.

Lemma 3 (Binary Search Lemma)
Let $\mathbf{x} = BinarySearch(\mathbf{z}, \mathbf{y}, \epsilon)$ defined as in Algorithm 2. Then one of 3 cases happen:

(i) $\mathbf{x} = \mathbf{y}$ and $\langle \boldsymbol{prox}(\mathbf{x}) - \mathbf{x}, \mathbf{x} - \mathbf{z} \rangle \geq 0$,
(ii) $\mathbf{x} = \mathbf{z}$ and $\langle \boldsymbol{prox}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \leq 0$, or
(iii) $\mathbf{x} = t\mathbf{y} + (1 - t)\mathbf{z}$ for some $t \in (0, 1)$ and $|\langle \boldsymbol{prox}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle| \leq 3\|\mathbf{y} - \mathbf{z}\|_2^2 \epsilon$.

Proof of Lemma 3 Items (i) and (ii), are simply Steps 2 and 5, respectively. For item (iii), we have

$$\begin{aligned}
&\|\mathbf{x} - \mathbf{w}\|_2 \\
&= \|t\mathbf{y} + (1 - t)\mathbf{z} - t^*\mathbf{y} - (1 - t^*)\mathbf{z}\|_2 \\
&= \|(t - t^*)\mathbf{y} - (t - t^*)\mathbf{z}\|_2 \\
&\leq \epsilon\|\mathbf{y} - \mathbf{z}\|_2.
\end{aligned}$$

Now it follows that

$$\begin{aligned}
&|\langle \mathbf{prox}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle| \\
&= |\langle \mathbf{prox}(\mathbf{x}) - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle - \langle \mathbf{prox}(\mathbf{w}) - \mathbf{w}, \mathbf{y} - \mathbf{z} \rangle| \\
&\leq \|\langle \mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{w}), \mathbf{y} - \mathbf{z} \rangle\|_2 + |\langle \mathbf{x} - \mathbf{w}, \mathbf{y} - \mathbf{z} \rangle| \\
&\leq \|\mathbf{prox}(\mathbf{x}) - \mathbf{prox}(\mathbf{w})\|_2 \|\mathbf{y} - \mathbf{z}\|_2 \\
&\quad + \|\mathbf{x} - \mathbf{w}\|_2 \|\mathbf{y} - \mathbf{z}\|_2 \\
&\leq 2\|\mathbf{x} - \mathbf{w}\|_2 \|\mathbf{y} - \mathbf{z}\|_2 \\
&\quad + \|\mathbf{x} - \mathbf{w}\|_2 \|\mathbf{y} - \mathbf{z}\|_2 \\
&= 3\|\mathbf{x} - \mathbf{w}\|_2 \|\mathbf{y} - \mathbf{z}\|_2 \\
&\leq 3\epsilon\|\mathbf{y} - \mathbf{z}\|_2^2.
\end{aligned}$$

Where the third inequality follows by Lemma 2 ∎

Using the above result, we can prove the following:

Corollary 1
Let $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{z}_k$ and ϵ_k be defined as in Algorithm 1 and $\eta_k L_k \geq 1$. Then for all $k \geq 1$,

$$\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle \leq (\eta_k L_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle + \frac{DL\eta_k L_k}{T^3}.$$

Proof of Corollary 1 Note that by Step 3 of Algorithm 1), $\mathbf{p}_k = -L(\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k)$. For $k = 1$, since $\mathbf{x}_1 = \mathbf{y}_1 = \mathbf{z}_1$, the inequality is trivially true. For $k \geq 2$, we consider the three cases of Lemma 3: (i) if $\mathbf{x}_k = \mathbf{y}_k$, the right hand side is $1/T \geq 0$ and the left hand side is $\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle = \langle -L(\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), \mathbf{x}_k - \mathbf{z}_k \rangle \leq 0$, (ii) if $\mathbf{x}_k = \mathbf{z}_k$, the left hand side

is 0 and $\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle = \langle -L(\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \rangle \geq 0$, so the inequality holds trivially, and (iii) in this last case, for some $t \in (0,1)$, we have

$$\begin{aligned}
&\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle \\
&= \langle -L(\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), t\mathbf{y}_k + (1-t)\mathbf{z}_k - \mathbf{z}_k \rangle \\
&= -Lt\langle (\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), \mathbf{y}_k - \mathbf{z}_k \rangle,
\end{aligned}$$

and

$$\begin{aligned}
&\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle \\
&= \langle -L(\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), \mathbf{y}_k - t\mathbf{y}_k - (1-t)\mathbf{z}_k \rangle \\
&= -L(1-t)\langle (\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), (\mathbf{y}_k - \mathbf{z}_k) \rangle.
\end{aligned}$$

Hence

$$\begin{aligned}
&\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle - (\eta_k L_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle \\
&\leq |\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle - (\eta_k L_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle| \\
&= |(-Lt + (\eta_k L_k - 1)L(1-t)) \\
&\qquad \langle (\mathbf{prox}(\mathbf{x}_k) - \mathbf{x}_k), (\mathbf{y}_k - \mathbf{z}_k) \rangle| \\
&\leq 3|(-Lt + (\eta_k L_k - 1)L(1-t))| \|\mathbf{y}_k - \mathbf{z}_k\|_2^2 \epsilon_k \\
&= 3|\eta_k L_k (1-t) + 1| L \|\mathbf{y}_k - \mathbf{z}_k\|_2^2 \epsilon_k \\
&= 3(\eta_k L_k + 1) L \|\mathbf{y}_k - \mathbf{z}_k\|_2^2 \epsilon_k \\
&= 6\eta_k L_k L \|\mathbf{y}_k - \mathbf{z}_k\|_2^2 \epsilon_k \\
&= \frac{6D\eta_k L_k L \|\mathbf{y}_k - \mathbf{z}_k\|_2^2}{D} \frac{1}{6dT^3} \\
&\leq \frac{DL\eta_k L_k}{T^3},
\end{aligned}$$

where in the last line we used the fact that $\|y_k - z_k\|_2^2 \leq Dd$ ∎

Similar to 1 for Algorithm 1, the following Lemma proves an analogous result for Algorithm 3.

Corollary 2
Let $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{z}_k$ and ϵ_k be defined as in Algorithm 3 and $\eta_k \tilde{L}_k \geq 1$. Then for all $k \geq 1$,

$$\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle \leq (\eta_k \tilde{L}_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle + \frac{DL\eta_k \tilde{L}_k}{T^3}.$$

Proof of Corollary 2 We consider two cases:

1. If $\mathbf{x}_k$ is generated through Algorithm 5, then $\mathbf{x}_k = \text{BinarySearch}(\mathbf{y}_k, \mathbf{z}_k, \epsilon)$ and $\tilde{L}_k = L_k$, so the statement follows from Corollary 1.

2. If x_k is generated through Algorithm 4, then $\mathbf{x}_k = \left(1 - \frac{1}{\eta_k \tilde{L}_k}\right)\mathbf{y}_k + \frac{1}{\eta_k \tilde{L}_k}\mathbf{z}_k$, and so satisfies

$$\langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k \rangle = (\eta_k \tilde{L}_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k \rangle$$

Next, we state a result regarding the mirror descent step. Similar results can be found in most texts on online optimization, e.g. [1].

Lemma 4 (Mirror Descent Inequality)
Let $\mathbf{z}_{k+1} = \arg\min_{\mathbf{z} \in \mathcal{C}} \langle \eta_k \mathbf{p}_k, \mathbf{z} - \mathbf{z}_k \rangle + \frac{1}{2}\|\mathbf{z} - \mathbf{z}_k\|_{S_k}^2$ and $D := \sup_{\mathbf{x},\mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_\infty^2$ be the diameter of $\mathcal{C}$ measured by infinity norm. Then for any $\mathbf{u} \in \mathcal{C}$, we have

$$\sum_{k=1}^{T} \langle \eta_k \mathbf{p}_k, \mathbf{z}_k - u \rangle \leq \sum_{k=1}^{T} \frac{\eta_k^2}{2} \|\mathbf{p}_k\|_{S_k^*}^2 + \frac{D}{2}\|\mathbf{s}_T\|_1$$

Proof of Lemma 4 For any $\mathbf{u} \in \mathcal{C}$ and by optimality of $\mathbf{z}_{k+1}$, we have $\langle \eta_k \mathbf{p}_k, \mathbf{z}_{k+1} - \mathbf{u} \rangle \leq \langle S_k(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{u} - \mathbf{z}_{k+1} \rangle$. Hence, using (5) and (4), it follows that

$$\begin{aligned}
&\langle \eta_k \mathbf{p}_k, \mathbf{z}_k - \mathbf{u} \rangle \\
&= \langle \eta_k \mathbf{p}_k, \mathbf{z}_k - \mathbf{z}_{k+1} \rangle + \langle \eta_k \mathbf{p}_k, \mathbf{z}_{k+1} - \mathbf{u} \rangle \\
&\leq \langle \eta_k \mathbf{p}_k, \mathbf{z}_k - \mathbf{z}_{k+1} \rangle - \langle S_k(\mathbf{z}_{k+1} - \mathbf{z}_k), \mathbf{z}_{k+1} - \mathbf{u} \rangle \\
&= \langle \eta_k \mathbf{p}_k, \mathbf{z}_k - \mathbf{z}_{k+1} \rangle - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|_{S_k}^2 \\
&\quad - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{u}\|_{S_k}^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{z}_k\|_{S_k}^2 \\
&\leq \sup_{\mathbf{z} \in \mathbb{R}^d} \left\{ \langle \eta_k \mathbf{p}_k, \mathbf{z} \rangle - \frac{1}{2}\|\mathbf{z}\|_{S_k}^2 \right\} \\
&\quad - \frac{1}{2}\|\mathbf{z}_{k+1} - \mathbf{u}\|_{S_k}^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{z}_k\|_{S_k}^2 \\
&= \frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 - \frac{1}{2}\|\mathbf{u} - \mathbf{z}_{k+1}\|_{S_k}^2 + \frac{1}{2}\|\mathbf{u} - \mathbf{z}_k\|_{S_k}^2.
\end{aligned}$$

Now recalling from Steps 5- 7 of Algorithm 1 that $S_k = \text{diag}(\mathbf{s}_k) + \delta\mathbb{I}$ and $\mathbf{s}_k \geq \mathbf{s}_{k-1}$, we sum over k to

get

$$
\begin{aligned}
&\sum_{k=1}^{T}\langle \eta_k \mathbf{p}_k, \mathbf{z}_k - u\rangle \\
&\le \sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 + \frac{1}{2}\|\mathbf{u}-\mathbf{z}_1\|_{S_1}^2 \\
&\quad + \sum_{k=2}^{T}\frac{1}{2}\|\mathbf{u}-\mathbf{z}_k\|_{S_k}^2 - \frac{1}{2}\|\mathbf{u}-\mathbf{z}_k\|_{S_{k-1}}^2 \\
&= \sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 + \frac{1}{2}\|\mathbf{u}-\mathbf{z}_1\|_{S_1}^2 \\
&\quad + \frac{1}{2}\sum_{k=2}^{T}\langle (S_k - S_{k-1})(\mathbf{u}-\mathbf{z}_k), \mathbf{u}-\mathbf{z}_k\rangle \\
&\le \sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 + \frac{1}{2}\|\mathbf{u}-\mathbf{z}_1\|_{\infty}^2\langle \mathbf{s}_1, \mathbf{1}\rangle \\
&\quad + \frac{1}{2}\sum_{k=2}^{T}\|\mathbf{u}-\mathbf{z}_k\|_{\infty}^2\langle \mathbf{s}_k - \mathbf{s}_{k-1}, \mathbf{1}\rangle \\
&\le \sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 + \frac{D}{2}\langle s_1, \mathbf{1}\rangle + \frac{D}{2}\sum_{k=2}^{T}\langle \mathbf{s}_k - \mathbf{s}_{k-1}, \mathbf{1}\rangle \\
&= \sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^*}^2 + \frac{D}{2}\|\mathbf{s}_T\|_1
\end{aligned}
$$

∎

Finally, we state a similar result to that of [17] that captures the benefits of using S_k in FLAG.

Lemma 5 (AdaGrad Inequalities) *Define $q_T := \sum_{i=1}^{d}\|G_T(i,:)\|_2$, where G_k is as in Step 5 of Algorithm 1. We have*

(i) $\sum_{k=1}^{T}\mathbf{g}_k^T S_k^{-1}\mathbf{g}_k \le 2q_T$,
(ii) $q_T^2 = \min_{S\in\mathcal{S}}\sum_{k=1}^{T}\mathbf{g}_k^T S^{-1}\mathbf{g}_k$, where $\mathcal{S} := \{S \in \mathbb{R}^{d\times d} \mid S \text{ is diagonal}, S_{ii} > 0, \text{trace}(S) \le 1\}$, and
(iii) $\sqrt{T} \le q_T \le \sqrt{dT}$.

Proof of Lemma 5 To prove part (i), we use the following inequality introduced in the proof of Lemma 4 in [17]: for any arbitrary real-valued sequence of $\{a_i\}_{i=1}^T$ and its vector representation as $a_{1:T} = [a_1, a_2, \ldots, a_T]$, we have

$$\sum_{k=1}^{T}\frac{a_k^2}{\|a_{1:k}\|_2} \le 2\|a_{1:T}\|_2.$$

So it follows that

$$
\begin{aligned}
&\sum_{k=1}^{T}\mathbf{g}_k^T S_k^{-1}\mathbf{g}_k \\
&= \sum_{k=1}^{T}\sum_{i=1}^{d}\frac{\mathbf{g}_k^2(i)}{\mathbf{s}_k^2(i)} \\
&= \sum_{i=1}^{d}\sum_{k=1}^{T}\frac{\mathbf{g}_k^2(i)}{\mathbf{s}_k(i)} \\
&= \sum_{i=1}^{d}\sum_{k=1}^{T}\frac{\mathbf{g}_k^2(i)}{\|G_k(i,:)\|_2} \\
&\le 2q_T,
\end{aligned}
$$

where the last equality follows from the definition of $\mathbf{s}_k$ in Step 6 of Algorithm 1.

For the rest of the proof, one can easily see that

$$\sum_{k=1}^{T}\mathbf{g}_k^T S^{-1}\mathbf{g}_k = \sum_{k=1}^{T}\sum_{i=1}^{d}\frac{\mathbf{g}_k^2(i)}{\mathbf{s}(i)} = \sum_{i=1}^{d}\frac{a(i)}{\mathbf{s}(i)},$$

where $a(i) := \sum_{k=1}^{T}\mathbf{g}_k^2(i)$ and $\mathbf{s} = \text{diag}(S)$. Now the Lagrangian for $\lambda \ge 0$ and $\boldsymbol{\nu} \ge \mathbf{0}$, can be written as

$$\mathcal{L}(\mathbf{s}, \lambda, \boldsymbol{\nu}) = \sum_{i=1}^{d}\frac{a(i)}{\mathbf{s}(i)} + \lambda\left(\sum_{i=1}^{d}\mathbf{s}(i) - 1\right) + \langle \boldsymbol{\nu}, \mathbf{s}\rangle.$$

Since the strong duality holds, for any primal-dual optimal solutions, S^*, λ^* and $\boldsymbol{\nu}^*$, it follows from complementary slackness that $\boldsymbol{\nu}^* = \mathbf{0}$ (since $\mathbf{s}^* > \mathbf{0}$). Now requiring that $\partial\mathcal{L}(\mathbf{s}^*, \lambda^*, \boldsymbol{\nu}^*)/\partial\mathbf{s}(i) = 0$ gives $\lambda^*\mathbf{s}^*(i) = \sqrt{a_i} > 0$, which since $\mathbf{s}^*(i) > 0$, implies that $\lambda^* > 0$. As a result, by using complementary slackness again, we must have $\sum_{i=1}^{d}\mathbf{s}^*(i) = 1$. Now simple algebraic calculations gives $\mathbf{s}^*(i) = \sqrt{a_i}/(\sum_{i=1}^{d}\sqrt{a_i})$ and part (ii) follows.

For part (iii), recall that $\|\mathbf{g}_k\|_2 = 1$. Now, since $\lambda_{\min}(S^{01}) \ge 1$, one has $1 \le \mathbf{g}_k^T S^{-1}\mathbf{g}_k$, and so $q_T \ge 1$. One the other hand, consider the optimization problem

$$
\begin{aligned}
\max\ &\sum_{i=1}^{d}\|G_T(i,:)\|_2 = \sum_{i=1}^{d}\sqrt{\sum_{k=1}^{T}\mathbf{g}_i^2(k)} \\
\text{s.t. }&\|\mathbf{g}_k\|_2^2 = 1,\ k = 1, 2, \ldots, T.
\end{aligned}
$$

The Lagrangian can be written as

$$
\begin{aligned}
\mathcal{L}(\{\mathbf{g}_k\}_{k=1}^T, \{\lambda\}_{k=1}^T) = &\sum_{i=1}^{d}\sqrt{\sum_{k=1}^{T}\mathbf{g}_i^2(k)} \\
&+ \sum_{k=1}^{T}\lambda_k\left(1 - \sum_{i=1}^{d}\mathbf{g}_i^2(k)\right).
\end{aligned}
$$

By KKT necessary condition, we require that $\partial\mathcal{L}(\{\mathbf{g}_k\}_{k=1}^T, \{\lambda\}_{k=1}^T)/\partial \mathbf{g}_i(k) = 0$, which implies that $\lambda_k = 1/(2\sqrt{\sum_{k=1}^T \mathbf{g}_i^2(k)})$, $i = 1, 2, \ldots, d$. Hence, $T = \sum_{i=1}^d \sum_{k=1}^T \mathbf{g}_i^2(k) = d/(4\lambda_k^2)$, and so $2\lambda_k = \sqrt{d/T}$, which gives $q_T \leq \sqrt{dT}$. ∎

We can now prove the central theorems of which is used to obtain FLAG's main result.

Theorem 3
Let $D := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{C}} \|\mathbf{x}-\mathbf{y}\|_\infty^2$. For any $\mathbf{u} \in \mathcal{C}$, after T iterations of Algorithm 1, we get

$$\sum_{k=1}^T \left\{ \left(\eta_{k-1}^2 L_{k-1} - \eta_k^2 L_k + \eta_k\right) F(\mathbf{y}_k) - \eta_k F(\mathbf{u}) \right\} + \eta_T^2 L_T F(\mathbf{y}_{T+1}) \leq \sum_{k=1}^T \frac{DL\eta_k^2 L_k}{T^3} + \frac{D}{2}\|\mathbf{s}_T\|_1.$$

Proof of Theorem 3 Noting that $\mathbf{p}_k = -L(\mathbf{y}_{k+1} - \mathbf{x}_k)$ is the gradient mapping of F on $\mathcal{C}$, it follows that

$$\begin{aligned}
&\sum_{k=1}^T \eta_k (F(\mathbf{y}_{k+1}) - F(\mathbf{u})) \\
&= \sum_{k=1}^T \eta_k (F(\mathbf{prox}(\mathbf{x}_k)) - F(\mathbf{u})) \\
&\leq \sum_{k=1}^T \eta_k \langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{u}\rangle - \frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2 \\
&= \sum_{k=1}^T \eta_k \langle \mathbf{p}_k, (\mathbf{z}_k - \mathbf{u})\rangle + \sum_{k=1}^T \eta_k \langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k\rangle - \frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2 \\
&\leq \sum_{k=1}^T \frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^{-1}}^2 + \frac{D}{2}\|\mathbf{s}_T\|_1 + \sum_{k=1}^T \eta_k \langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k\rangle - \frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2 \\
&= \sum_{k=1}^T \frac{\eta_k(\eta_k L_k - 1)}{2L}\|\mathbf{p}_k\|_2^2 + \frac{D}{2}\|\mathbf{s}_T\|_1 + \sum_{k=1}^T \eta_k \langle \mathbf{p}_k, \mathbf{x}_k - \mathbf{z}_k\rangle \\
&\leq \sum_{k=1}^T \frac{\eta_k(\eta_k L_k - 1)}{2L}\|\mathbf{p}_k\|_2^2 + \frac{D}{2}\|\mathbf{s}_T\|_1 \\
&\qquad + \sum_{k=1}^T \left(\eta_k(\eta_k L_k - 1)\langle \mathbf{p}_k, \mathbf{y}_k - \mathbf{x}_k\rangle + \frac{DL\eta_k^2 L_k}{T^3} \right) \\
&\leq \sum_{k=1}^T \frac{DL\eta_k^2 L_k}{T^3} + \frac{D}{2}\|\mathbf{s}_T\|_1 \\
&\quad + \sum_{k=1}^T \eta_k(\eta_k L_k - 1)\left(F(\mathbf{y}_k) - F(\mathbf{y}_{k+1})\right). \quad \text{(Lemma 1)}
\end{aligned}$$

Where the first inequality is by Lemma 1, the second inequality is by Lemma 4, the third equality is by Step 8 of Algorithm 1, and the second last inequality is by Corollary 1. Now we have

$$\begin{aligned}
&\sum_{k=1}^T \eta_k (F(\mathbf{y}_{k+1}) - F(\mathbf{u})) - \eta_k(\eta_k L_k - 1)\left(F(\mathbf{y}_k) - F(\mathbf{y}_{k+1})\right) \\
&= \sum_{k=1}^T \eta_k F(\mathbf{y}_{k+1}) - \eta_k F(\mathbf{u}) - \eta_k(\eta_k L_k - 1)F(\mathbf{y}_k) \\
&\quad + \eta_k(\eta_k L_k - 1)F(\mathbf{y}_{k+1}) \\
&= \sum_{k=1}^T \eta_k^2 L_k F(\mathbf{y}_{k+1}) - \eta_k F(\mathbf{u}) - \eta_k(\eta_k L_k - 1)F(\mathbf{y}_k) \\
&= \eta_T^2 L_T F(\mathbf{y}_{T+1}) \\
&\quad + \sum_{k=1}^T \eta_{k-1}^2 L_{k-1} F(\mathbf{y}_k) - \eta_k F(\mathbf{u}) - \eta_k(\eta_k L_k - 1)F(\mathbf{y}_k) \\
&= \eta_T^2 L_T F(\mathbf{y}_{T+1}) \\
&\quad + \sum_{k=1}^T \left(\eta_{k-1}^2 L_{k-1} - \eta_k^2 L_k + \eta_k\right) F(\mathbf{y}_k) - \eta_k F(\mathbf{u}),
\end{aligned}$$

and the result follows. ∎

Once again, we present the analog of Theorem 3 for Algorithm 3.

Theorem 4
Let $D := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{C}} \|\mathbf{x}-\mathbf{y}\|_\infty^2$. For any $\mathbf{u}\in\mathcal{C}$, after T iterations of Algorithm 1, we get

$$\begin{aligned}
&\sum_{k=1}^{T}\left\{\left(\eta_{k-1}^2\tilde{L}_{k-1}-\eta_k^2\tilde{L}_k+\eta_k\right)F(\mathbf{y}_k)-\eta_k F(\mathbf{u})\right\}\\
&\quad+\eta_T^2\tilde{L}_T F(\mathbf{y}_{T+1})\\
&\leq\sum_{k=1}^{T}\frac{D\tilde{L}\eta_k^2\tilde{L}_k}{T^3}+\frac{D}{2}\|\mathbf{s}_T\|_1.
\end{aligned}$$

Proof of Theorem 4 Parts of this proof which differ from the proof of Theorem 3 are bolded. Noting that $\mathbf{p}_k=-L(\mathbf{y}_{k+1}-\mathbf{x}_k)$ is the gradient mapping of F on $\mathcal{C}$, it follows that

$$\begin{aligned}
&\sum_{k=1}^{T}\eta_k(F(\mathbf{y}_{k+1})-F(\mathbf{u}))\\
&=\sum_{k=1}^{T}\eta_k(F(\mathbf{prox}(\mathbf{x}_k))-F(\mathbf{u}))\\
&\leq\sum_{k=1}^{T}\eta_k\langle\mathbf{p}_k,\mathbf{x}_k-\mathbf{u}\rangle-\frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2\\
&=\sum_{k=1}^{T}\eta_k\langle\mathbf{p}_k,(\mathbf{z}_k-\mathbf{u})\rangle+\sum_{k=1}^{T}\eta_k\langle\mathbf{p}_k,\mathbf{x}_k-\mathbf{z}_k\rangle\\
&\quad-\frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2\\
&\leq\sum_{k=1}^{T}\frac{\eta_k^2}{2}\|\mathbf{p}_k\|_{S_k^{-1}}^2+\frac{D}{2}\|\mathbf{s}_T\|_1+\sum_{k=1}^{T}\eta_k\langle\mathbf{p}_k,\mathbf{x}_k-\mathbf{z}_k\rangle\\
&\quad-\frac{\eta_k}{2L}\|\mathbf{p}_k\|_2^2\\
&=\sum_{k=1}^{T}\frac{\eta_k(\eta_k\tilde{L}_k-1)}{2L}\|\mathbf{p}_k\|_2^2+\frac{D}{2}\|\mathbf{s}_T\|_1\\
&\qquad+\sum_{k=1}^{T}\eta_k\langle\mathbf{p}_k,\mathbf{x}_k-\mathbf{z}_k\rangle\\
&\leq\sum_{k=1}^{T}\frac{\eta_k(\eta_k\tilde{L}_k-1)}{2L}\|\mathbf{p}_k\|_2^2+\frac{D}{2}\|\mathbf{s}_T\|_1\\
&\quad+\sum_{k=1}^{T}\left(\eta_k(\eta_k\tilde{L}_k-1)\langle\mathbf{p}_k,\mathbf{y}_k-\mathbf{x}_k\rangle+\frac{DL\eta_k^2\tilde{L}_k}{T^3}\right)\\
&\leq\sum_{k=1}^{T}\frac{DL\eta_k^2\tilde{L}_k}{T^3}+\frac{D}{2}\|\mathbf{s}_T\|_1\\
&\quad+\sum_{k=1}^{T}\eta_k(\eta_k\tilde{L}_k-1)\left(F(\mathbf{y}_k)-F(\mathbf{y}_{k+1})\right).
\end{aligned}$$

Where the first inequality follows from Lemma 1, the second inequality follows from Lemma 4, the last equality follows from Steps 9 and 11 of Alg 4, Steps 8 and 9 of Alg 5, and the second last inequality follows from Corollary 2, and the last equality follows from Lemma 1.

Now we have

$$\begin{aligned}
&\sum_{k=1}^{T} \eta_k (F(\mathbf{y}_{k+1}) - F(\mathbf{u})) \\
&\quad - \eta_k(\eta_k \tilde{L}_k - 1)\left(F(\mathbf{y}_k) - F(\mathbf{y}_{k+1})\right) \\
&= \sum_{k=1}^{T} \eta_k F(\mathbf{y}_{k+1}) - \eta_k F(\mathbf{u}) - \eta_k(\eta_k \tilde{L}_k - 1) F(\mathbf{y}_k) \\
&\quad + \eta_k(\eta_k \tilde{L}_k - 1) F(\mathbf{y}_{k+1}) \\
&= \sum_{k=1}^{T} \eta_k^2 L_k F(\mathbf{y}_{k+1}) - \eta_k F(\mathbf{u}) - \eta_k(\eta_k \tilde{L}_k - 1) F(\mathbf{y}_k) \\
&= \eta_T^2 \tilde{L}_T F(\mathbf{y}_{T+1}) \\
&\quad + \sum_{k=1}^{T} \eta_{k-1}^2 \tilde{L}_{k-1} F(\mathbf{y}_k) - \eta_k F(\mathbf{u}) \\
&\quad - \eta_k(\eta_k \tilde{L}_k - 1) F(\mathbf{y}_k) \\
&= \eta_T^2 \tilde{L}_T F(\mathbf{y}_{T+1}) \\
&\quad + \sum_{k=1}^{T} \left(\eta_{k-1}^2 \tilde{L}_{k-1} - \eta_k^2 \tilde{L}_k + \eta_k\right) F(\mathbf{y}_k) - \eta_k F(\mathbf{u}),
\end{aligned}$$

and the result follows. ∎

We now set out to put the final piece of the proof in place: choosing the stepsize η_k for the mirror descent step.

Lemma 6
For the choice of η_k in Algorithm 1 and $k \geq 1$, we have

(i) $\eta_k^2 L_k = \sum_{i=1}^{k} \eta_i$,
(ii) $\eta_{k-1}^2 L_{k-1} - \eta_k^2 L_k + \eta_k = 0$, *and*
(iii) $\eta_k L_k \geq 1$.

Proof We prove (i) by induction. For $k = 1$, is is easy to verify that $\eta_1 = 1/L_1$, and so $\eta_1^2 L_1 = \eta_1$ and the base case follows trivially. Now suppose $\eta_{k-1}^2 L_{k-1} = \sum_{i=1}^{k-1} \eta_i$. Re-arranging (i) for k gives

$$0 = \eta_k^2 L_k - \eta_k - \sum_{i=1}^{k-1} \eta_i = \eta_k^2 L_k - \eta_k - \eta_{k-1}^2 L_{k-1}.$$

Now, it is easy to verify that the choice of η_k in Algorithm 1 is a solution of the above quadratic equation. The rest of the items follow immediately from part (i). ∎

Once again, the FLARE analog of Lemma 6 is

Lemma 7
For the choice of η_k in Algorithm 3 and $k \geq 1$, we have

(i) $\eta_k^2 \tilde{L}_k = \sum_{i=1}^{k} \eta_i$,
(ii) $\eta_{k-1}^2 \tilde{L}_{k-1} - \eta_k^2 \tilde{L}_k + \eta_k = 0$, *and*
(iii) $\eta_k \tilde{L}_k \geq 1$.

Proof of Lemma 7 Completely identical to proof of Lemma 6.

Corollary 3
Let $D := \sup_{\mathbf{x},\mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_\infty^2$. For any $\mathbf{u} \in \mathcal{C}$, after T iterations of Algorithm 1, we get

$$F(\mathbf{y}_{T+1}) - F(\mathbf{u}) \leq \frac{LD}{T^2} + \frac{D\|\mathbf{s}_T\|_1}{2\sum_{k=1}^{T} \eta_k}.$$

Proof of corollary 3 The result follows from Theorem 3 and Lemma 6 as well as noting that $\eta_k^2 L_k = \sum_{i=1}^{k} \eta_i \leq \sum_{i=1}^{T} \eta_i = \eta_T^2 L_T$. ∎

The FLARE analog:

Corollary 4
Let $D := \sup_{\mathbf{x},\mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_\infty^2$. For any $\mathbf{u} \in \mathcal{C}$, after T iterations of Algorithm 3, we get

$$F(\mathbf{y}_{T+1}) - F(\mathbf{u}) \leq \frac{LD}{T^2} + \frac{D\|\mathbf{s}_T\|_1}{2\sum_{k=1}^{T} \eta_k}.$$

Proof of corollary 4 The result follows from Theorem 4 and Lemma 7 as well as noting that $\eta_k^2 L_k = \sum_{i=1}^{k} \eta_i \leq \sum_{i=1}^{T} \eta_i = \eta_T^2 \tilde{L}_T$. ∎

Finally, it only remains to lower bound $\sum_{k=1}^{T} \eta_k$, which is done in the following Lemma.

Lemma 8
For the choice of η_k in Algorithm 1, we have

$$\sum_{k=1}^{T} \eta_k \geq \frac{T^3}{1000 \sum_{k=1}^{T} L_k}$$

Proof of Lemma 8 We prove by induction on T. For $T = 1$, we have $\eta_1 = 1/L_1$, and the base case holds trivially. Suppose the desired relation holds for $T-1$. We have

$$\begin{aligned}\sum_{k=1}^{T}\eta_k &= \sum_{k=1}^{T-1}\eta_k + \eta_T \\ &\geq \frac{(T-1)^3}{1000\sum_{k=1}^{T-1}L_k} + \frac{1}{2L_T} \\ &+ \sqrt{\frac{1}{4L_T^2} + \frac{(T-1)^3}{1000L_T\sum_{k=1}^{T-1}L_k}} \\ &\geq \frac{(T-1)^3}{1000\sum_{k=1}^{T-1}L_k} + \sqrt{\frac{(T-1)^3}{1000L_T\sum_{k=1}^{T-1}L_k}} \\ &\geq \frac{(T-1)^3}{1000\sum_{k=1}^{T-1}L_k} + \sqrt{\frac{T^3}{8000L_T\sum_{k=1}^{T-1}L_k}}.\end{aligned}$$

Where the first inequality is by the induction hypothesis on η_k. Now if

$$\frac{(T-1)^3}{1000\sum_{k=1}^{T-1}L_k} \geq \frac{T^3}{1000\sum_{k=1}^{T}L_k},$$

then we are done. Otherwise denoting $\alpha := \sum_{k=1}^{T}L_k$, we must have that

$$\begin{aligned}L_T &\leq \frac{\alpha T^3 - \alpha(T-1)^3}{T^3} \\ &= \frac{\alpha T^3 - \alpha\left(T^3 - 3T^2 + 3T - 1\right)}{T^3} \\ &= \frac{\alpha(3T^2 - 3T + 1)}{T^3} \\ &\leq \frac{4\sum_{k=1}^{T}L_k}{T}.\end{aligned}$$

Hence, we get

$$\begin{aligned}\sum_{k=1}^{T}\eta_k &\geq \frac{(T-1)^3}{1000\sum_{k=1}^{T-1}L_k} + \sqrt{\frac{T^4}{32000L_T\left(\sum_{k=1}^{T}L_k\right)^2}} \\ &\geq \frac{(T-1)^3}{1000\sum_{k=1}^{T}L_k} + \frac{4T^2}{1000\sum_{k=1}^{T}L_k} \\ &\geq \frac{T^3}{1000\sum_{k=1}^{T}L_k}.\end{aligned}$$

∎

Remark: We note here that we made little effort to minimize constants, and that we used rather sloppy bounds such as $T - 1 \geq T/2$. As a result, the constant appearing above is very conservative and a mere by product of our proof technique.

Lemma 9
For the choice of η_k in Algorithm 3, we have

$$\sum_{k=1}^{T}\eta_k \geq \frac{T^3}{\lambda \cdot 1000\sum_{k=1}^{T}L_k}$$

Proof of Lemma 9 Once again, exactly identical to the proof of Lemma 8, we have

$$\sum_{k=1}^{T}\eta_k \geq \frac{T^3}{1000\sum_{k=1}^{T}\tilde{L}_k}$$

Finally, using the guarantee that $\tilde{L}_k \leq \lambda L_k$ from Step 11 of Algorithm 4 and Step 9 from Algorithm 5, we get the conclusion.

The proof of FLAG's main result, Theorem 1, follows rather immediately.

Proof of Theorem 1 The result follows immediately from Lemma 8 and Corollary 3 and noting that $\sum_{k=1}^{T}L_k = L\sum_{k=1}^{T}\mathbf{g}_k^T S_k^{-1}\mathbf{g}_k \leq 2Lq_T$ by Lemma 5 and $\|\mathbf{s}_T\|_1 = q_T$ by Step 6 of Algorithm 1 and definition of q_T in Lemma 5. This gives

$$F(\mathbf{y}_{T+1}) - F(\mathbf{u}) \leq \frac{LD}{T^2} + \frac{q_T^2}{T}\frac{1000LD}{T^2} \leq \frac{q_T^2}{T}\frac{1001LD}{T^2}.$$

Now from Lemma 5, we see that $\beta := q_T^2/T \in [1, d]$. Finally, the run-time per iteration follows from having to do $\log_2(1/\epsilon)$ calls to bisection, each taking $\mathcal{O}(\mathcal{T}_{\mathbf{prox}})$ time. ∎

The proof of FLARE's main result, Theorem 2, is obtained similarly to that of Theorem 1.

Proof of Theorem 2 The result follows immediately from Lemma 9 and Corollary 4 and noting that $\sum_{k=1}^{T}L_k = L\sum_{k=1}^{T}\mathbf{g}_k^T S_k^{-1}\mathbf{g}_k \leq 2Lq_T$ by Lemma 5 and $\|\mathbf{s}_T\|_1 = q_T$ by Step 6 of Algorithm 4 and Step 5 of Algorithm 5 and definition of q_T in Lemma 5. This gives

$$\begin{aligned}F(\mathbf{y}_{T+1}) - F(\mathbf{u}) &\leq \frac{LD}{T^2} + \frac{q_T^2}{T}\frac{1000\lambda LD}{T^2} \\ &\leq \frac{q_T^2}{T}\frac{1001\lambda LD}{T^2}.\end{aligned}$$

Now from Lemma 5, we see that $\beta := q_T^2/T \in [1, d]$. Finally, we try to guess a suitable $\tilde{L}_k$ for $\log(d/\epsilon)$ times, and resort to BinarySearch after. If we resort

to algorithm 5 (essentially BinarySeaerch), we make $\log(1/\epsilon)$ calls to bisection, so overall the number of inner iterations per outer iteration is same as Algorithm 1. Each inner iteration takes $\mathcal{O}(\mathcal{T}_{\mathbf{prox}})$ time in the worst case (if we have to resort to algorithm 5 each time). ∎