APPENDIX: PROOFS

Proof of Lemma 1

We first prove a technical lemma that bounds the \( \ell_\infty \) norm of error vectors.

**Lemma 4.** For any \( x \in \mathbb{R}^d \) and \( z_i \in \{\pm 1\}^d \), with probability \( 1 - O(d^{-3}) \) (conditioned on \( x_t \) and \( z_i \))

\[
\left\| \sum_{i=1}^{n} \varepsilon_i z_i \right\|_\infty \lesssim \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + H\delta.
\]

*Proof.* Let \( \tilde{\xi}_i = \xi_i/\delta \sim \mathcal{N}(0, \sigma^2/\delta^2) \). Consider the following decomposition:

\[
\left\| \sum_{i=1}^{n} \varepsilon_i z_i \right\|_\infty \leq \frac{1}{n\delta} \left\| \sum_{i=1}^{n} \tilde{\xi}_i z_i \right\|_\infty + \delta \cdot \sup_{1 \leq i \leq n} \left| z_i^\top H_i(\kappa_i, z_i) \right| \cdot \|z_i\|_\infty.
\]

The second term on the right-hand side of the above inequality is upper bounded by \( O(H\delta) \) almost surely, because \( \|z_i\|_\infty \leq 1 \) and \( |z_i^\top H_i(\kappa_i, z_i)| \leq \|H_i(\kappa_i, z_i)\|_1 \cdot \|z_i\|_\infty \leq H \). For the first term, because \( \tilde{\xi}_i \) are centered sub-Gaussian random variables independent of \( z_i \) and \( \|z_i\|_\infty \leq 1 \), we have that \( 1/n \cdot \sum_{i=1}^{n} \tilde{\xi}_i z_i \|_\infty \lesssim \sqrt{\sigma^2 \log d/n} \) with probability \( 1 - O(d^{-3}) \), by invoking standard sub-Gaussian concentration inequalities. \qed

Now define \( \hat{\theta} = (\hat{y}_1, \hat{\mu}_1) \), \( \theta_0 = (y_0, \delta^{-1} f(x_t)) \) and \( \bar{Z} = (\tilde{z}_1, \ldots, \tilde{z}_n) \) where \( \tilde{z}_i = (z, 1) \in \mathbb{R}^{d+1} \). Define also that \( Y = (\bar{y}_1, \ldots, \bar{y}_n) \). The estimator can then be written as \( \hat{\theta} = \arg \min_{\theta \in \mathbb{R}^{d+1}} \frac{1}{2} \| \bar{Y} - \bar{Z} \theta \|_2^2 + \lambda \| \theta \|_1 \) where \( \bar{Y} = \bar{Z} \theta_0 + \varepsilon \), \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \). We first establish a “basic inequality” type results that are essential in performance analysis of Lasso type estimators. By optimality of \( \hat{\theta} \), we have that

\[
\frac{1}{n} \| Y - \bar{Z} \hat{\theta} \|_2^2 + \lambda \| \hat{\theta} \|_1 \leq \frac{1}{n} \| Y - \bar{Z} \theta_0 \|_2^2 + \lambda \| \theta_0 \|_1 = \frac{1}{n} \| \varepsilon \|_2^2 + \lambda \| \theta_0 \|_1.
\]

Re-organizing terms we obtain

\[
\lambda \| \hat{\theta} \|_1 \leq \lambda \| \theta_0 \|_1 + \frac{2}{n} \left( \| \hat{\theta} - \theta_0 \|_2 \bar{Z} \| \varepsilon \right) \| \varepsilon \|_2.
\]

On the other hand, by Hölder’s inequality and Lemma 4 we have, with probability \( 1 - O(d^{-2}) \),

\[
\frac{2}{n} \left( \| \hat{\theta} - \theta_0 \|_2 \bar{Z} \| \varepsilon \right) \| \varepsilon \|_2 \leq 2 \| \hat{\theta} - \theta_0 \|_1 \cdot \left( \frac{1}{n} \| \bar{Z} \|_\infty \| \varepsilon \|_\infty \right) \lesssim \| \hat{\theta} - \theta_0 \|_1 \cdot \left( \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + H\delta \right).
\]

Subsequently, if \( \lambda \leq c_0 (\sigma \delta^{-1} \sqrt{\log d/n} + H\delta) \) for some sufficiently small \( c_0 > 0 \), we have that \( \| \hat{\theta} \|_1 \leq \| \theta_0 \|_1 + 1/2 \| \hat{\theta} - \theta_0 \|_1 \). Multiplying by 2 and adding \( \| \hat{\theta} - \theta_0 \|_1 \) on both sides of the inequality we obtain \( \| \hat{\theta} - \theta_0 \|_1 \leq 2(\| \hat{\theta} - \theta_0 \|_1 + \| \theta_0 \|_1 - \| \hat{\theta} \|_1) \). Recall that \( \theta_0 \) is sparse and let \( \bar{S} = S \cup \{d + 1\} \) be the support of \( \theta_0 \). We then have \( \| (\hat{\theta} - \theta_0)_{\bar{S}} \|_1 + \| (\theta_0)_{\bar{S}} \|_1 - \| \theta_{\bar{S}} \|_1 = 0 \) and hence \( \| (\hat{\theta} - \theta_0)_{\bar{S}} \|_1 - \| (\theta - \theta_0)_{\bar{S}} \|_1 \leq \| \hat{\theta} - \theta_0 \|_1 \leq 2(\| \hat{\theta} - \theta_0 \|_S \|_{\bar{S}} \|_1 \). Thus,

\[
\| (\hat{\theta} - \theta_0)_{\bar{S}} \|_1 \leq 3(\| \hat{\theta} - \theta_0 \|_{\bar{S}} \|_{\bar{S}} \|_1.
\]

Now consider \( \hat{\theta} \) that minimizes \( \frac{1}{n} \| Y - \bar{Z} \theta \|_2^2 + \lambda \| \theta \|_1 \). By KKT condition we have that

\[
\left\| \frac{1}{n} \bar{Z}^\top (Y - \bar{Z} \theta) \right\|_\infty \leq \frac{\lambda}{2}.
\]

Define \( \hat{\Sigma} = \frac{1}{n} \bar{Z}^\top \bar{Z} \) and recall that \( Y = \bar{Z} \theta_0 + \varepsilon \). Invoking Lemma 4 and the scaling of \( \lambda \) we have that, with probability \( 1 - O(d^{-2}) \),

\[
\| \hat{\Sigma}(\hat{\theta} - \theta_0) \|_\infty \leq \frac{\lambda}{2} + \left\| \frac{1}{n} \bar{Z}^\top \varepsilon \right\|_\infty \lesssim \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + \delta H.
\]
By definition of \( \{ z_i \}_{i=1}^n \), we know that \( \bar{\Sigma}_{jj} = 1 \) for all \( j = 1, \ldots, d+1 \) and \( E[\bar{\Sigma}_{jk}] = 0 \) for \( j \neq k \). By Hoeffding’s inequality [17] and union bound we have that with probability \( 1 - O(d^{-2}) \), \( \| \bar{\Sigma} - I_{(d+1) \times (d+1)} \|_\infty \lesssim \sqrt{\log d/n} \), where \( \| \cdot \|_\infty \) denotes the maximum absolute value of matrix entries. Also note that \( \theta - \theta_0 \) satisfies \( \| (\theta - \theta_0)_{\mathbb{S}} \|_1 \leq 3\| (\theta - \theta_0)_{\mathbb{S}} \|_1 \) thanks to Eq. (8). Subsequently,

\[
\| \hat{\theta} - \theta_0 \|_\infty \leq \| \hat{\Sigma}(\hat{\theta} - \theta_0) \|_\infty + \| \hat{\Sigma} - I \|_\infty \| \hat{\theta} - \theta_0 \|_1 \\
\leq \| \hat{\Sigma}(\hat{\theta} - \theta_0) \|_\infty + \| \bar{\Sigma} - I \|_\infty \| \hat{\theta} - \theta_0 \|_1 \\
\leq \| \bar{\Sigma}(\hat{\theta} - \theta_0) \|_\infty + \| \bar{\Sigma} - I \|_\infty \cdot 4\| \hat{\theta} - \theta_0 \|_2 \\
\leq \| \bar{\Sigma}(\hat{\theta} - \theta_0) \|_\infty + \| \bar{\Sigma} - I \|_\infty \cdot 4(s+1)\| \hat{\theta} - \theta_0 \|_\infty \\
\leq \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + \delta H + \sqrt{\frac{s^2 \log d}{n}} \cdot \| \hat{\theta} - \theta_0 \|_\infty. \tag{10}
\]

Combining Eq. (10) together with the scaling \( n = \Omega(s^2 \log d) \) we complete the proof of Lemma 1. Note that the statement on the \( \ell_1 \) error \( \| \theta - \theta_0 \|_1 \) is a simple consequence of the basic inequality Eq. (8).

**Proof of Theorem 1**

The basis of our algorithm is the analysis of the finite-difference algorithm proposed by [13] under low dimensions. In particular, applying the analysis in [2] for low-dimensional strongly smooth functions, we have for every epoch \( t < s \)

\[
E[f(x_t)] - \inf_{x \in \bar{\mathcal{X}}, x_{\bar{S}} = 0} f(x) \lesssim \text{poly}(s, \sigma, H, \| x^*_0 \|_1) \cdot T^{-1/3},
\]

where \( x_t \) is the solution point at the \( t \)th epoch in Algorithm 2 and \( \text{poly}(\cdot) \) is any polynomial function of constant degrees. Recall that \( \| x^*_0 \|_1 \leq \| x^* \|_1 \leq B \) by Assumption (A2). Using Markov’s inequality we have that with probability 0.9,

\[
f(x_t) - \inf_{x \in \bar{\mathcal{X}}, x_{\bar{S}} = 0} f(x) \lesssim \text{poly}(s, \sigma, H, \| x^*_0 \|_1) \cdot T^{-1/3}, \quad \forall t = 0, \ldots, s. \tag{11}
\]

We are now ready to prove Theorem 1. Let \( \hat{S} = \hat{S}_t \) be the subset when Algorithm 2 terminates. In the rest of the proof we assume the conclusions in Corollary 1 and Lemma 1 hold, which happens with probability \( 1 - O(d^{-1}) \). Define \( \Delta \hat{S} = \hat{S} \setminus \overline{S} \), \( x^* := \inf_{x \in \bar{\mathcal{X}}} f(x) \) and \( x^*_t := \inf_{x \in \bar{\mathcal{X}}, x_{\bar{S}} = 0} f(x) \). Assumption (A5) implies that \( x^* \) can be chosen such that \( x^*_{\mathbb{S}} = 0 \). Also, if \( \Delta \hat{S} = \emptyset \) we know that \( x^*_t = x^* \) and Theorem 1 automatically holds due to Eq. (11). Therefore in the rest of the proof we shall assume that \( \Delta \hat{S} \neq \emptyset \).

Because \( \Delta \hat{S} \neq \emptyset \) and \( |S| = s \), we must have \( |\hat{S}_t| < s \). From the description of Algorithm 2, it can only happen with \( \hat{S}_t = \hat{S}_{t-1} \). We then have that

\[
f(x_{T+1}) - f(x^*) = f(x^*_{t-1}) - f(x^*) + f(\bar{x}_{t-1}) - f(x^*_{t-1}) \\
\leq f(x^*_{t-1}) - f(x^*) + \text{poly}(s, \sigma, H, \| x^* \|_1) \cdot T^{-1/3} \\
\leq \nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) + \text{poly}(s, \sigma, H, \| x^* \|_1) \cdot T^{-1/3}, \tag{12}
\]

where Eq. (12) holds with probability at least 0.9, thanks to Eq. (11). Because \( x^*_{t-1} \) is the minimizer of \( f \) on vectors in \( \bar{\mathcal{X}} \) that are supported on \( \hat{S} = \hat{S}_{t-1} = \hat{S}_t \), and that both \( x^*_{t-1} \) and \( x^* \) truncated on \( \hat{S} \) are feasible (i.e., in the restrained set \( \bar{\mathcal{X}} \)), it must hold that \( \| \nabla f(x^*_{t-1}) \|_{\Delta \hat{S}} \leq 0 \) by first-order optimality conditions. On the other hand, by Corollary 1 and the definition of \( \hat{S}_t \), we have that \( \| \nabla f(x^*_{t-1}) \|_{\Delta \hat{S}} \| \lesssim 2\eta \). Also note that \( (x^* - x^*_{t-1})_{\mathbb{S}} = 0 \) and \( |x^*_{t-1} - x^*_{\mathbb{S}}| = 0 \). Subsequently,

\[
\nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) \leq \| \nabla f(x^*_{t-1}) \|_{\Delta \hat{S}} \leq \| \nabla f(x^*_{t-1}) \|_{\Delta \hat{S}} \| x^*_{\mathbb{S}} \|_1 \leq 2\eta \|
\]

Combining Eqs. (13,14) and the scalings of \( \eta, \delta, \lambda \) and \( T' = T/2s \) we complete the proof of Theorem 1.
Proof of Lemma 2

We use the “full-length” parameterization \( \hat{\theta}_t = \tilde{\theta}_t + \frac{1}{n} Z_t^\top (\tilde{Y}_t - Z_t \tilde{\theta}_t) \), where \( \tilde{\theta}_t, Z_t \) and \( \tilde{Y}_t \) are notations defined in the proof of Lemma 1 (with subscripts \( t \) added to emphasize that both \( Z_t \) and \( \tilde{Y}_t \) are specific to the \( t \)th epoch in Algorithm 3). Because \( \tilde{Y}_t = Z_t \theta_{0t} + \varepsilon_t \) (where \( \theta_{0t} = \nabla f(x_t) \) and \( \varepsilon = (\varepsilon_{t1}, \ldots, \varepsilon_{tn}) \), with \( \varepsilon_{ti} \) defined in Eq. (2)). we have

\[
\tilde{\theta}_t = \hat{\theta}_t + \frac{1}{n} \sum_{i=1}^{n} Z_t^\top (Z_t \theta_{0t} + \varepsilon_t - \tilde{Z}_t \tilde{\theta}_t) = \theta_{0t} + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_t + (\tilde{\Sigma} - I_{(d+1) \times (d+1)}) (\hat{\theta}_t - \theta_{0t}),
\]

where \( \tilde{\Sigma} = \frac{1}{n} \tilde{Z}_t^\top \tilde{Z}_t \). Recall that \( \varepsilon_{ti} = \xi_i / \delta + \delta z_i^\top H_i (\kappa_i, z_i) z_i \). Define \( b_i = z_i^\top H_i (\kappa_i, z_i) z_i \) and \( b = (b_1, \ldots, b_n) \).

Also note that the first \( d \) components of \( \tilde{\theta}_t \) are identical to \( g_t \) defined in Eq. (5). Subsequently,

\[
\tilde{g}_t = g_t + \frac{1}{n \delta} \sum_{i=1}^{n} Z_t^\top (Z_t \theta_{0t} + \varepsilon_t - \tilde{Z}_t \tilde{\theta}_t) = \theta_{0t} + \frac{1}{n \delta} \sum_{i=1}^{n} \varepsilon_t + (\tilde{\Sigma} - I_{(d+1) \times (d+1)}) (\hat{\theta}_t - \theta_{0t}),
\]

In Eq. (15) we divide \( \tilde{g}_t - g_t \) into two terms. We first consider the term \( \zeta_t := \frac{1}{n \delta} \sum_{i=1}^{n} Z_t^\top \xi_i \). It is clear that \( E[\zeta_t | x_t] = 0 \) because \( E[\xi_t | x_t, Z_t] = 0 \). Now consider any \( d \)-dimensional vector \( a \in \mathbb{R}^d \), and to simplify notations all derivations below are conditioned on \( x_t \). For any \( i \in [n] \), \( z_{ti} a \) are i.i.d. sub-Gaussian random variables with common parameter \( \nu^2 = ||a||_2^2 \). Also, \( \xi_t \) is a sub-Gaussian random variable with parameter \( \sigma^2 \) and is independent of \( z_{ti} a \). Thus, invoking Lemma 6 we have that \( \zeta_t z_{ti} a \) is a sub-exponential random variable with parameters \( \nu = \alpha / \sqrt{2} \leq \sigma ||a||_2 \). Consequently, \( \langle \zeta_t, a \rangle = \frac{1}{n \delta} \sum_{i=1}^{n} \xi_t z_{ti} a \) is a centered sub-exponential random variable with parameters \( \nu = \sqrt{n \delta / 2} \cdot \alpha \leq \sigma ||a||_2 / \sqrt{n} \).

We next consider the term \( \gamma_t = \frac{1}{n \delta} \sum_{i=1}^{n} \xi_t z_{ti} a + (\tilde{\Sigma} - I) (\hat{\theta}_t - \theta_{0t}) \). By Assumption (A3) we know that \( ||b||_\infty \leq \delta H \).

Subsequently, by Hölder’s inequality we have that

\[
||\gamma_t|| \leq \frac{\delta}{n} ||Z_t||_{1, \infty} ||b||_{\infty} + ||\tilde{\Sigma} - I||_{\infty} ||\hat{\theta}_t - \theta_{0t}||_1 \leq H \delta + \sqrt{\frac{\log d}{n}} \left( \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n} + s \delta H} \right).
\]

where the second inequality holds with probability \( 1 - O(d^{-2}) \) thanks to Lemma 1.

Proof of Theorem 2

We first note that the cumulative regret \( R^C_A(T) \) can be upper bounded as

\[
R^C_A(T) \leq \left[ \frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f^* \right] + \sup_t \sup_{z \in \{-1,1\}^d} \left| f(x_t + \delta z) - f(x_t) \right|.
\]

Because \( || \nabla f(x) ||_1 \leq H \) for all \( x \in \mathcal{X} \) and \( z \in \{-1,1\}^d \), using Hölder’s inequality we have that

\[
\left| f(x_t + \delta z) - f(x_t) \right| \leq \delta H \leq B \left( \frac{s \log^2 d}{T} \right)^{1/4},
\]

which is a second-order term. Thus, to prove upper bounds on \( R^C_A(T) \) it suffices to consider only \( \frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f^* \).

We next cite the result in [22] that gives explicit cumulative regret bounds for mirror descent with approximate gradients:

**Lemma 5** ([22], Lemma 3). Let \( || \cdot \|_\psi \) and \( || \cdot \|_{\psi'} \) be a pair of conjugate norms, and let \( \Delta_\psi(\cdot, \cdot) \) be a Bregman divergence that is \( \kappa \)-strongly convex with respect to \( || \cdot \|_\psi \). Suppose \( f \) is \( H \)-smooth with respect to \( || \cdot \|_\psi \), meaning
that \( f(y) \leq f(x) + \nabla f(x)^	op (y - x) + \frac{\tilde{H}}{2} \|x - y\|_\psi^2 \) for all \( x, y \in \mathcal{X} \), and \( \eta < \kappa/\tilde{H} \). Define \( g_t = \nabla f(x_t) \), and let \( x_0, \ldots, x_{T-1} \) be iterations in Algorithm 3. Then for every \( 0 \leq t \leq T' - 1 \) and any \( x^* \in \mathcal{X} \),

\[
\eta [f(x_{t+1}) - f(x^*)] + \Delta_\psi(x_{t+1}, x^*) \leq \Delta_\psi(x_t, x^*) + \eta \langle \tilde{g}_t - g_t, x^* - x_t \rangle + \eta \frac{\|\tilde{g}_t - g_t\|_\psi^2}{2(\kappa - H \eta)}. \tag{16}
\]

Adding both sides of Eq. (16) from \( t = 0 \) to \( t = T' - 1 \), telescoping and noting that \( \Delta_\psi(x_{T'}, x^*) \geq 0 \), we obtain

\[
\frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) \leq \frac{\Delta_\psi(x_0, x^*)}{\eta T'} + \frac{1}{T'} \sum_{t=0}^{T'-1} \langle \tilde{g}_t - g_t, x_t - x^* \rangle + \frac{\eta}{2(\kappa - H \eta)} \sup_{0 \leq t < T'} \|\tilde{g}_t - g_t\|_\psi^2. \tag{17}
\]

Set \( \| \cdot \| = \| \cdot \|_a \) for \( a = \frac{2 \log d}{2 \log d - 1} \). It is easy to verify that under Assumption (A3), the function \( f \) satisfies

\[
f(y) \geq f(x) + \nabla f(x)^	op (y - x) + H \|y - x\|_\infty^2,
\]

\[
\geq f(x) + \nabla f(x)^	op (y - x) + \tilde{H} \|y - x\|_\psi^2
\]

for all \( x, y \in \mathcal{X} \) with \( \tilde{H} \leq eH \), because \( \|y - x\|_\psi^2 \leq d^{(1-1/a)} \|y - x\|_a^2 \leq d^{1/\log d} \|y - x\|_1^2 = e \|y - x\|_1^2 \) by Hölder’s inequality. In addition, by definition of Bregman divergence we have that

\[
\Delta_\psi(x_0, x^*) \leq \frac{1}{2(a - 1)} \|x^\ast\|_\infty^2 \leq \frac{1}{2(a - 1)} \|x^\ast\|_2^2 \leq \|x^\ast\|_2^2 \log d \leq B^2 \log d,
\]

(18)

where the first inequality holds because \( \psi_a(x_0) = \psi_a(0) = 0 \) and \( \nabla \psi_a(x_0) = \nabla \psi_a(0) = 0 \) for \( a > 1 \).

We next upper bound the \( \frac{1}{T'} \sum_{t=0}^{T'-1} \langle \tilde{g}_t - g_t, x^* - x_t \rangle \) and \( \|\tilde{g}_t - g_t\|_\psi^2 \), terms. By Lemma 2 and sub-exponential concentration inequalities (e.g., Lemma 7), we have that with probability \( 1 - O(d^{-1}) \)

\[
\|\tilde{g}_t - g_t\|_\psi^2 \leq \|\zeta_t\|_\infty + \|\gamma_t\|_\infty \lesssim \frac{\sigma}{\delta} \left( \frac{\log d}{T} + \frac{\log d}{n} \right) + H \delta + \frac{\sigma s \log d}{\delta n} \lesssim \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + H \delta
\]

uniformly over all \( t' \in \{0, \ldots, T'-1\} \), the last inequality holds because \( n = \Omega(s^2 \log d) \). Subsequently, by Hölder’s inequality we have that

\[
\sup_{0 \leq t < T'} \|\tilde{g}_t - g_t\|_\psi^2 \lesssim d^{(2a-1)/a} \sup_{0 \leq t < T'} \|\tilde{g}_t - g_t\|_2^2 \lesssim \frac{\sigma^2 \log d}{\delta^2 n} + H^2 \sigma^2. \tag{19}
\]

We now consider the first term \( \frac{1}{T'} \sum_{t=0}^{T'-1} \langle \zeta_t, x^* - x_t \rangle \leq \frac{1}{T'} \sum_{t=0}^{T'-1} X_t + \sup_{0 \leq t < T'-1} \|\gamma^\ast\|_\infty \|x^\ast - x_t\|_1 \), where \( X_t := \langle \zeta_t, x^\ast - x_t \rangle \). By Lemma 2, we know that \( X_1, X_1, \ldots, X_{T-1} \) is a centered sub-exponential random variable with parameters \( \nu = \sqrt{n}/2 \cdot \alpha \lesssim \sigma \|x^\ast - x_t\|_2/\delta \sqrt{n} \lesssim \sigma \|x^\ast\|_1/\delta \sqrt{n} \). Invoking concentration inequalities for sub-exponential martingales ([40], also phrased as Lemma 8 for a simplified version in the appendix) and the definition that \( T' = T/n \), we have with probability \( 1 - O(d^{-1}) \)

\[
\left| \frac{1}{T'} \sum_{t=0}^{T'-1} \langle \zeta_t, x^* - x_t \rangle \right| \lesssim \frac{\sigma \|x^\ast\|_1}{\delta} \left( \sqrt{\frac{\log d}{T'}} + \frac{\log d}{T} \right) \lesssim \frac{\sigma \|x^\ast\|_1}{\delta} \sqrt{\frac{\log d}{T}},
\]

where the last inequality holds because \( T \geq n = \Omega(s^2 \log d) \). Thus,

\[
\left| \frac{1}{T'} \sum_{t=0}^{T'-1} \langle \tilde{g}_t - g_t, x^* - x_t \rangle \right| \lesssim \frac{\sigma \|x^\ast\|_1}{\delta} \sqrt{\frac{\log d}{T'}} + \frac{\log d}{T} \left( H \delta + \frac{\sigma s \log d}{\delta n} \right). \tag{20}
\]

Combining Eqs. (18,19,20) with Eq. (17) and taking \( x^\ast \) to be a minimizer of \( f \) on \( \mathcal{X} \) that satisfies \( \|x^\ast\|_1 \leq B \), we obtain

\[
\frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) \lesssim \frac{\|x^\ast\|_1^2 \log d}{\eta T} + \frac{\sigma \|x^\ast\|_1}{\delta} \sqrt{\frac{\log d}{T'}} + \frac{\|x^\ast\|_1}{T} \left( H \delta + \frac{\sigma s \log d}{\delta n} \right) + \eta \left( \frac{\sigma^2 \log d}{\delta^2 n} + H^2 \sigma^2 \right)
\]
\[ \frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) \leq B \left( \frac{n \log d}{T} + \frac{\sigma B}{\sqrt{n}} + \frac{B(\sigma + H)}{\sqrt{T}} \right) \]

with probability \( 1 - \mathcal{O}(d^{-1}) \), provided that \( \eta < \kappa / 2H = 1/2H \).

We are now ready to prove Theorem 2. By the conditions we impose on \( T \) and the choices of \( \eta \) and \( n \), it is easy to verify that \( \eta < 1/2H, n = \Omega(s^2 \log d) \) and \( n = \mathcal{O}(T) \). Subsequently,

\[
\begin{align*}
\frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) &\leq B \left( \frac{n \log d}{T} + \frac{\sigma B}{\sqrt{n}} + \frac{B(\sigma + H)}{\sqrt{T}} \right) \\
&\leq B \left( \frac{(1 + H)^2 \log d}{T} \right)^{1/4} + \frac{\sigma B}{\sqrt{T}} \left( \frac{s \log d}{T} \right)^{1/4} + \frac{B(\sigma + H)}{\sqrt{T}} \left( \frac{s \log d}{T} \right)^{1/4} + \mathcal{O}(T^{-1/2}) \\
&\leq \left( B \sqrt{\log d} + \frac{\sigma B}{\sqrt{s}} + \frac{\sigma^2 B}{s} \right)^{1/4} + B(\sigma + H) \sqrt{\log d} \left( \frac{s \log d}{T} \right)^{1/4} + \mathcal{O}(T^{-1/2}) \\
&\leq (1 + \sigma + \sigma^2 / s) B \sqrt{\log d} \left( \frac{(1 + H)^2 s}{T} \right)^{1/4} + \mathcal{O}(T^{-1/2}).
\end{align*}
\]

Proof of Lemma 3

Using the model Eq. (2) we can decompose \( \tilde{g}_t(\delta) - g_t \) as

\[
\begin{align*}
\tilde{g}_t(\delta) - g_t &= \frac{\delta}{2} \mathbb{E} \left[ (z^T H_t z) \right] + \frac{\delta}{2n} \sum_{i=1}^{n} (z_i^T H_t z_i) z_i - \mathbb{E} \left[ (z^T H_t z) \right] \\
&+ \frac{\delta}{2n} \sum_{i=1}^{n} (z_i^T (H_t (\delta z_i) - H_t) z_i) z_i + \left[ \left( \tilde{\Sigma} - I \right) (\tilde{\theta}_t - \theta_t) \right]_{1:d},
\end{align*}
\]

where \( \tilde{\Sigma}, \tilde{\theta}_t \) and \( \theta_t \) are similarly defined as in the proof of Lemma 2. The sub-exponentiality of \( \langle \tilde{\xi}_t(\delta), a \rangle \) for any \( a \in \mathbb{R}^d \) is established in Lemma 2. We next consider \( \tilde{\beta}_t(\delta) \). For any \( a \in \mathbb{R}^d \) consider \( \tilde{\beta}_t(\delta), a = \sum_{i=1}^{n} X_i(a) \) where \( X_i(a) = (z_i^T H_t z_i)(z_i^T a) - \mathbb{E} [(z_i^T H_t z_i)(z_i^T a)] \) are centered i.i.d. random variables conditioned on \( H_t \) and \( x_t \). In addition, \( |X_i(a)| \leq 2 \|H_t\|_1 \|z_i\|_{\infty} \cdot \|a\|_1 \|z_i\|_{\infty} \leq H \|a\|_1 \) almost surely. Therefore, \( X_i(a) \) is a sub-Gaussian random variable with parameter \( \nu = H \|a\|_1 \), and hence \( \langle \tilde{\beta}_t(\delta), a \rangle \) is a sub-Gaussian random variable with parameter \( \nu = \delta H \|a\|_1 / \sqrt{n} \). Finally, for the deterministic term \( \gamma_t(\delta) \), we have that

\[
\|\gamma_t(\delta)\|_\infty \leq \frac{\delta}{2} \sup_{\delta_{z} \in \{\pm 1\}^d} \|H_t(\delta z) - H_t\|_1 \|z\|_\infty^2 + ||(\tilde{\Sigma} - I)(\tilde{\theta}_t - \theta_t)\|_\infty
\]

\[
\leq \frac{\delta}{2} \sup_{\|\delta z\|_\infty \leq L \cdot \|\delta z\|_\infty} \|\delta z\|_\infty \|z\|_\infty^2 + ||(\tilde{\Sigma} - I)\|_{\text{max}} \|\tilde{\theta}_t - \theta_t\|_{\infty}
\]

\[
\leq L \delta^2 + \sqrt{\frac{\log d}{n}} \left( \frac{\sigma}{\delta} \sqrt{\frac{\log d}{n}} + s \delta H \right)
\]

\[
\leq L \delta^2 + \frac{\sigma s \log d}{n \delta} + s \delta H \sqrt{\frac{\log d}{n}}.
\]

Proof of Theorem 3

Because \( f \) is convex, \( R_{\mathcal{A}}(T) = f(x_{T+1}) - f^* \leq \frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f^* \). Thus it suffices to upper bound \( \frac{1}{T} \sum_{t=0}^{T-1} f(x_t) - f(x^*) \), where \( x^* \in \mathcal{X}, \|x^*\|_1 \leq B \) is a minimizer of \( f \) over \( \mathcal{X} \). Using the strategy in the proof of Theorem 2, this amounts to upper bound (with high probability) \( \|\tilde{g}_t(w) - g_t\|_{\psi^*}^2 \) and \( \frac{1}{T} \sum_{t=0}^{T-1} (\tilde{g}_t(w) - g_t, x^* - x_t) \).
For the first term, using sub-exponentiality of $\tilde{\zeta}_t$ and sub-gaussianity of $\tilde{\beta}_t$, we have with probability $1 - O(d^{-1})$ uniformly over all $t \in \{0, \ldots, T' - 1\}$,
\[
\|\tilde{g}_t^w - g_t\|_\infty \leq \|\tilde{\zeta}_t\|_\infty + \|\tilde{\beta}_t\|_\infty + |\tilde{\gamma}_t|_\infty \\
\lesssim \frac{\sigma}{\delta} \left( \sqrt{\frac{\log d}{n} + \frac{\log d}{n}} \right) + \delta H \sqrt{\frac{\log d}{n} + L \delta^2} + H \delta \sqrt{\frac{s^2 \log d}{n} + \frac{\sigma s \log d}{\delta n}} \\
\lesssim \left( \frac{\sigma}{\delta} + s \delta H \right) \sqrt{\frac{\log d}{n} + L \delta^2},
\]
where the last inequality holds because $n = \Omega(s^2 \log d)$. Subsequently, with probability $1 - O(d^{-1})$
\[
\sup_{0 \leq t \leq T' - 1} \|\tilde{g}_t^w - g_t\|^2_{\psi^s} \lesssim \left( \frac{\sigma^2}{\delta^2} + s^2 \delta^2 H^2 \right) \frac{\log d}{n} + L^2 \delta^4.
\] (22)

For the other term $\frac{1}{T'} \sum_{t=0}^{T'-1} \langle \tilde{g}_t^w - g_t, x^* - x_t \rangle$, again using concentration inequalities of sub-exponential/sub-Gaussian martingales and noting that $\|x^* - x_t\|_2 \leq \|x^* - x_t\|_1 \leq 2B$, we have
\[
\frac{1}{T'} \sum_{t=0}^{T'-1} \langle \tilde{g}_t^w - g_t, x^* - x_t \rangle = \frac{1}{T'} \sum_{t=0}^{T'-1} (\tilde{\zeta}_t + \tilde{\beta}_t + \tilde{\gamma}_t, x^* - x_t) \\
\lesssim \left( \frac{\sigma}{\delta} + s \delta H \right) B \sqrt{\frac{\log d}{T}} + B \left( L \delta^2 + \frac{\sigma s \log d}{\delta n} + s \delta H \sqrt{\frac{\log d}{n}} \right).
\] (23)

Subsequently, combining Eqs. (22,23) with Eq. (17) we have
\[
\frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) \lesssim \frac{B^2 \log d}{\eta} \frac{n}{T} + \left( \frac{\sigma}{\delta} + s \delta H \right) \frac{\log d}{T} + \left( B + \eta \right) \left( L \delta^2 + \frac{\sigma s \log d}{\delta n} + s \delta H \sqrt{\frac{\log d}{n}} \right) \\
+ \eta \left( \frac{\sigma^2}{\delta^2} + s^2 \delta^2 H^2 \right) \frac{\log d}{n} + \eta L^2 \delta^4.
\] (24)

We are now ready to prove Theorem 3. It is easy to verify that with the condition imposed on $T$ and the selection of $\eta$ and $n$, it holds that $\eta < 1/2H$, $n = \Omega(s^2 \log d)$ and $n \leq T/10$. Subsequently,
\[
\frac{1}{T'} \sum_{t=0}^{T'-1} f(x_t) - f(x^*) \\
\lesssim B n^{1/3} \sqrt{\frac{\log d}{T}} + \left[ \sigma \left( \frac{n}{s \log d} \right)^{1/3} + \tilde{O}(n^{-1/3}) \right] B \sqrt{\frac{\log d}{T}} + \left( B + \tilde{O} \left( \frac{n^{2/3}}{\sqrt{T}} \right) \right) \left[ (L + \sigma) \left( \frac{s \log d}{n} \right)^{2/3} + \tilde{O}(n^{-5/6}) \right] \\
+ B n^{2/3} \sqrt{\frac{\log d}{T}} \left( \frac{s^2 \log d}{n} \right)^{2/3} + \tilde{O}(n^{-2/3}) \frac{\log d}{n} + B n^{2/3} \sqrt{\frac{\log d}{T}} \left( \frac{s^2 \log d}{n} \right)^{4/3} \\
\lesssim B n^{1/3} \sqrt{\frac{\log d}{T}} + \sigma B \left( \frac{n}{s \log d} \right)^{1/3} \sqrt{\frac{\log d}{T}} + B (L + \sigma) \left( \frac{s \log d}{n} \right)^{2/3} + \sigma^2 B \left( \frac{n}{s^2 \log^2 d} \right)^{1/3} \sqrt{\frac{\log d}{T}} + \tilde{O}(T^{-5/12}) \\
\lesssim \left( B \sqrt{\frac{\log d}{T}} + \frac{\sigma B \sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B \sqrt{\log d}}{s^{2/3}} \right) \left( \frac{(1 + L)s^{2/3}}{T} \right)^{1/3} + B (L + \sigma) \left( \frac{s^{2/3} \log d}{T} \right)^{1/3} + \tilde{O}(T^{-5/12}) \\
\lesssim \left( B \sqrt{\frac{\log d}{T}} + \frac{\sigma B \sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B \sqrt{\log d}}{s^{2/3}} \right) \left( \frac{(1 + L)s^{2/3}}{T} \right)^{1/3} + B \sigma \sqrt{\log d} \left( \frac{(1 + L)s^{2/3}}{T} \right)^{1/3} + \tilde{O}(T^{-5/12}) \\
\lesssim (1 + \sigma + \sigma^2 / s^{2/3}) B \sqrt{\log d} \left( \frac{(1 + L)s^{2/3}}{T} \right)^{1/3} + \tilde{O}(T^{-5/12}).
Additional tail inequalities

Lemma 6. Suppose $X$ and $Y$ are centered sub-Gaussian random variables with parameters $\nu_1^2$ and $\nu_2^2$, respectively. Then $XY$ is a centered sub-exponential random variable with parameter $\nu = \sqrt{2}v$ and $\alpha = 2v$, where $v = 2e^{2/e} + 1 \nu_1 \nu_2$.

Proof. $XY$ is clearly centered because $\mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y = 0$, thanks to independence. We next bound $\mathbb{E}(|XY|^k)$ for $k \geq 3$ (i.e., verification of the Bernstein’s condition). Because $X$ and $Y$ are independent, we have that $\mathbb{E}(|XY|^k) = \mathbb{E}|X|^k \cdot \mathbb{E}|Y|^k$. In addition, because $X$ is a centered sub-Gaussian random variable with parameter $\nu_1^2$, it holds that $(\mathbb{E}|X|^k)^{1/k} \leq \nu_1 e^{1/e} \sqrt{k}$. Similarly, $(\mathbb{E}|X|^k)^{1/k} \leq \nu_2 e^{1/e} \sqrt{k}$. Subsequently,

$$
\mathbb{E}|XY|^k \leq \left(e^{2/e} \nu_1 \nu_2\right)^k \cdot \left(e^{2/e} \nu_1 \nu_2\right)^k \cdot e^{k!} \leq \frac{1}{2} k! \cdot \left(2e^{2/e} + 1 \nu_1 \nu_2\right)^k .
$$

where in the second inequality we use the Stirling’s approximation inequality that $\sqrt{2\pi}k^{k}e^{-k} \leq k!$. The sub-exponential parameter of $XY$ can then be determined.

Lemma 7 (Bernstein’s inequality). Suppose $X$ is a sub-exponential random variable with parameters $\nu$ and $\alpha$.

$$
\Pr[|X - \mathbb{E}X| > t] \leq \begin{cases} 
2 \exp\left\{-t^2/2\nu^2\right\} , & 0 < t \leq \nu^2/\alpha; \\
2 \exp\left\{-t/2\alpha\right\} , & t > \nu^2/\alpha.
\end{cases}
$$

The following lemma is a simplified version of Theorem 1.2A in [40] (note that the original form in [40] is one-sided; the two-sided version below can be trivially obtained by considering $-X_1, \ldots, -X_n$ and applying the union bound).

Lemma 8 (Bernstein’s inequality for martingales). Suppose $X_1, \ldots, X_n$ are random variables such that $\mathbb{E}[X_j | X_1, \ldots, X_{j-1}] = 0$ and $\mathbb{E}[X_j^2 | X_1, \ldots, X_{j-1}] \leq \sigma^2$ for all $t = 1, \ldots, n$. Further assume that $\mathbb{E}|X_j|^k | X_1, \ldots, X_{j-1} | \leq \frac{1}{2} k! k^2 \sigma^{2k-2}$ for all integers $k \geq 3$. Then for all $t > 0$,

$$
\Pr\left[ \sum_{j=1}^n X_j \geq t \right] \leq 2 \exp \left\{ - \frac{t^2}{2(n\sigma^2 + bt)} \right\} .
$$