APPENDIX: PROOFS

Proof of Lemma 1

We first prove a technical lemma that bounds the ℓ_∞ norm of error vectors.

Lemma 4. *For any $x \in \mathbb{R}^d$ and $z_i \in \{\pm 1\}^d$, with probability $1-\mathcal{O}(d^{-3})$ (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 $\bar\xi_i = \xi_i/\delta \sim \mathcal{N}(0, \sigma^2/\delta^2)$. Consider the following decomposition:

$$\left\|\sum_{i=1}^n \varepsilon_i z_i\right\|_\infty \leq \frac{1}{n\delta}\left\|\sum_{i=1}^n \bar\xi_i z_i\right\|_\infty + \delta\cdot\sup_{1\leq i\leq n}\left|z_i^\top H_t(\kappa_i, z_i) z_i\right|\cdot\|z_i\|_\infty.$$

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

Now define $\widehat\theta = (\widehat g_t, \widehat\mu_t)$, $\theta_0 = (g_t, \delta^{-1}f(x_t))$ and $\bar Z = (\bar z_1, \ldots, \bar z_n)$ where $\bar z_i = (z_i, 1) \in \mathbb{R}^{d+1}$. Define also that $Y = (\widetilde y_1, \ldots, \widetilde y_n)$. The estimator can then be written as $\widehat\theta = \arg\min_{\theta\in\mathbb{R}^{d+1}} \frac{1}{n}\|\widetilde Y - \bar Z\theta\|_2^2 + \lambda\|\theta\|_1$ where $\widetilde Y = \bar Z\theta_0 + \varepsilon$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$. We first establish a "basic inequality" type results that are essential in performance analysis of Lasso type estimators. By optimality of $\widehat\theta$, we have that

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

Re-organizing terms we obtain

$$\lambda\|\widehat\theta\|_1 \leq \lambda\|\theta_0\|_1 + \frac{2}{n}(\widehat\theta - \theta_0)^\top \bar Z^\top\varepsilon.$$

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

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

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 $\|\widehat\theta\|_1 \leq \|\theta_0\|_1 + 1/2\|\widehat\theta - \theta_0\|_1$. Multiplying by 2 and adding $\|\widehat\theta - \theta_0\|_1$ on both sides of the inequality we obtain $\|\widehat\theta-\theta_0\|_1 \leq 2(\|\widehat\theta - \theta_0\|_1 + \|\widehat\theta_0\|_1 - \|\widehat\theta\|_1)$. Recall that θ_0 is sparse and let $\bar S = S\cup\{d+1\}$ be the support of θ_0. We then have $\|(\widehat\theta-\theta_0)_{\bar S^c} + \|(\theta_0)_{\bar S^c}\|_1 - \|\widehat\theta_{\bar S^c}\|_1 = 0$ and hence $\|(\widehat\theta-\theta_0)_{\bar S^c}\|_1 - \|(\widehat\theta-\theta_0)_{\bar S}\|_1 \leq \|\widehat\theta-\theta_0\|_1 \leq 2\|(\widehat\theta-\theta_0)_{\bar S}\|_1$. Thus,

$$\|(\widehat\theta-\theta_0)_{\bar S^c}\|_1 \leq 3\|(\widehat\theta-\theta_0)_{\bar S}\|_1. \tag{8}$$

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

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

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

$$\|\widehat\Sigma(\widehat\theta-\theta_0)\|_\infty \leq \frac{\lambda}{2} + \left\|\frac{1}{n}\bar Z^\top\varepsilon\right\| \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + \delta H. \tag{9}$$

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

$$\begin{aligned}
\|\widehat{\theta} - \theta_0\|_\infty &\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|(\widehat{\Sigma} - I)(\widehat{\theta} - \theta_0)\|_\infty \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \|\widehat{\theta} - \theta_0\|_1 \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \cdot 4\|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1 \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \cdot 4(s+1)\|\widehat{\theta} - \theta_0\|_\infty \\
&\lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + \delta H + \sqrt{\frac{s^2 \log d}{n}} \cdot \|\widehat{\theta} - \theta_0\|_\infty. \qquad (10)
\end{aligned}$$

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 ℓ_1 error $\|\widehat{\theta} - \theta_0\|_1$ 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 $t < s$

$$\mathbb{E}[f(x_t)] - \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x) \lesssim \mathrm{poly}(s, \sigma, H, \|x^*_{\widehat{S}_t}\|_1) \cdot T^{-1/3},$$

where x_t is the solution point at the tth epoch in Algorithm 2 and $\mathrm{poly}(\cdot)$ is any polynomial function of constant degrees. Recall that $\|x^*_{\widehat{S}_t}\|_1 \leq \|x^*\|_1 \leq B$ by Assumption (A2). Using Markov's inequality we have that with probability 0.9,

$$f(x_t) - \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x) \lesssim \mathrm{poly}(s, \sigma, H, \|x^*_{\widehat{S}_t}\|_1) \cdot T^{-1/3}, \qquad \forall t = 0, \ldots, s. \qquad (11)$$

We are now ready to prove Theorem 1. Let $\widehat{S} = \widehat{S}_t$ 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 $1 - \mathcal{O}(d^{-1})$. Define $\Delta S = S \backslash \widehat{S}$, $x^* := \inf_{x \in \mathcal{X}} f(x)$ and $x^*_t = \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x)$. Assumption (A5) implies that x^* can be chosen such that $x^*_{S^c} = 0$. Also, if $\Delta_S = \emptyset$ 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 $\Delta_S \neq \emptyset$.

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

$$\begin{aligned}
f(x_{T+1}) - f(x^*) &= f(x^*_{t-1}) - f(x^*) + f(\widehat{x}_{t-1}) - f(x^*_{t-1}) \\
&\leq f(x^*_{t-1}) - f(x^*) + \mathrm{poly}(s, \sigma, H, \|x^*\|_1) \cdot T^{-1/3} \qquad (12) \\
&\leq \nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) + \mathrm{poly}(s, \sigma, H, \|x^*\|_1) \cdot T^{-1/3}, \qquad (13)
\end{aligned}$$

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 $\widetilde{\mathcal{X}}$ that are supported on $\widehat{S} = \widehat{S}_{t-1} = \widehat{S}_t$, and that both x^*_{t-1} and x^* truncated on $\widehat{S}$ are feasible (i.e., in the restrained set $\widetilde{\mathcal{X}}$), it must hold that $\langle [\nabla f(x^*_{t-1})]_{\widehat{S}}, (x^*_{t-1} - x^*)_{\widehat{S}} \rangle \leq 0$ by first-order optimality conditions. On the other hand, by Corollary 1 and the definition of $\widehat{S}_t$, we have that $\|[\nabla f(x^*_{t-1})_{\Delta_S}]\|_\infty \leq 2\eta$. Also note that $(x^* - x^*_{t-1})_{S^c} = 0$ and $[x^*_{t-1}]_{\Delta_S} = 0$. Subsequently,

$$\nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) \leq \left| \langle \nabla f(x^*_{t-1})_{\Delta_S}, x^*_{\Delta_S} \rangle \right| \leq \|[\nabla f(x^*_{t-1})]_{\Delta_S}\|_\infty \|x^*_{\Delta_S}\|_1 \leq 2\eta \|x^*\|_1. \qquad (14)$$

Combining Eqs. (13,14) and the scalings of η, δ, λ and $T' = T/2s$ we complete the proof of Theorem 1.

Proof of Lemma 2

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

$$\widetilde{\theta}_t = \widehat{\theta}_t + \frac{1}{n}\bar{Z}_t^\top(\bar{Z}_t\theta_{0t} + \varepsilon_t - \bar{Z}_t\widehat{\theta}_t) = \theta_{0t} + \frac{1}{n}\bar{Z}_t^\top\varepsilon_t + (\widehat{\Sigma} - I_{(d+1)\times(d+1)})(\widehat{\theta}_t - \theta_{0t}),$$

where $\widehat{\Sigma} = \frac{1}{n}\bar{Z}_t^\top\bar{Z}_t$. Recall that $\varepsilon_{ti} = \xi_i/\delta + \delta z_i^\top H_t(\kappa_i, z_i)z_i$. Define $b_i = z_i^\top H_t(\kappa_i, z_i)z_i$ and $b = (b_1, \ldots, b_n)$. Also note that the first d components of $\widetilde{\theta}_t$ are identical to $\widetilde{g}_t$ defined in Eq. (5). Subsequently,

$$\widehat{g}_t = g_t + \underbrace{\frac{1}{n\delta}Z_t^\top\xi}_{:=\zeta_t} + \underbrace{\frac{\delta}{n}Z_t^\top b + \Big[(\widehat{\Sigma} - I_{(d+1)\times(d+1)})(\widehat{\theta}_t - \theta_{0t})\Big]_{1:d}}_{:=\gamma_t}. \tag{15}$$

In Eq. (15) we divide $\widehat{g}_t - g_t$ into two terms. We first consider the term $\zeta_t := \frac{1}{n\delta}Z_t^\top\xi$. It is clear that $\mathbb{E}[\zeta_t|x_t] = 0$ because $\mathbb{E}[\xi|x_t, Z_t] = 0$. 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 $i \in [n]$, $z_{ti}^\top a$ are i.i.d. sub-Gaussian random variables with common parameter $\nu^2 = \|a\|_2^2$. Also, ξ_i is a sub-Gaussian random variable with parameter σ^2 and is independent of $z_{ti}^\top a$. Thus, invoking Lemma 6 we have that $\xi_i z_{ti}^\top a$ is a sub-exponential random variable with parameters $\nu = \alpha/\sqrt{2} \lesssim \sigma\|a\|_2$. Consequently, $\langle\zeta_t, a\rangle = \frac{1}{n\delta}\sum_{i=1}^n \xi_i z_{ti}^\top a$ is a centered sub-exponential random variable with parameters $\nu = \sqrt{n/2}\cdot\alpha \lesssim \sigma\|a\|_2/\delta\sqrt{n}$.

We next consider the term $\gamma_t = \frac{\delta}{n}Z_t^\top b + (\widehat{\Sigma} - I)(\widehat{\theta}_t - \theta_{0t})$. By Assumption (A3) we know that $\|b\|_\infty \le \delta H$. Subsequently, by Hölder's inequality we have that

$$\begin{aligned}
\|\gamma_t\|_\infty &\le \frac{\delta}{n}\|Z_t\|_{1,\infty}\|b\|_\infty + \|\widehat{\Sigma} - I\|_\infty\|\widehat{\theta}_t - \theta_{t0}\|_1 \\
&\lesssim H\delta + \sqrt{\frac{\log d}{n}}\left(\frac{\sigma s}{\delta}\sqrt{\frac{\log d}{n}} + s\delta H\right).
\end{aligned}$$

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

Proof of Theorem 2

We first note that the cumulative regret $R_{\mathcal{A}}^{\mathsf{C}}(T)$ can be upper bounded as

$$R_{\mathcal{A}}^{\mathsf{C}}(T) \lesssim \left[\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*\right] + \sup_t \sup_{z\in\{\pm1\}^d}\big|f(x_t + \delta z) - f(x_t)\big|.$$

Because $\|\nabla f(x)\|_1 \le H$ for all $x \in \mathcal{X}$ and $z \in \{\pm1\}^d$, using Hölder's inequality we have that

$$\big|f(x_t + \delta z) - f(x_t)\big| \le \delta H \lesssim B\left(\frac{s\log^2 d}{T}\right)^{1/4},$$

which is a second-order term. Thus, to prove upper bounds on $R_{\mathcal{A}}^{\mathsf{C}}(T)$ 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 κ-strongly convex with respect to $\|\cdot\|_\psi$. Suppose f is $\widetilde{H}$-smooth with respect to $\|\cdot\|_\psi$, meaning* *that* $f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{\widetilde{H}}{2}\|x-y\|_\psi^2$ *for all* $x, y \in \mathcal{X}$, *and* $\eta < \kappa/\widetilde{H}$. *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* $x^* \in \widetilde{\mathcal{X}}$,

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

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

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

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

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^\top (y-x) + H\|y-x\|_\infty^2 \\ &\geq f(x) + \nabla f(x)^\top (y-x) + \widetilde{H}\|y-x\|_\psi^2 \end{aligned}$$

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

$$\Delta_\psi(x_0, x^*) \leq \frac{1}{2(a-1)}\|x^*\|_a^2 \leq \frac{1}{2(a-1)}\|x^*\|_1^2 \leq \|x^*\|_1^2 \log d \leq B^2 \log d, \tag{18}$$

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

$$\|\widetilde{g}_t - g_t\|_\infty \leq \|\zeta_t\|_\infty + \|\gamma_t\|_\infty \lesssim \frac{\sigma}{\delta}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right) + H\delta + \frac{\sigma s \log d}{\delta n} \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + H\delta$$

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

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

We now consider the first term $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t - g_t, x^* - x_t\rangle \leq \frac{1}{T'}\sum_{t=0}^{T'-1} X_t + \sup_{0 \leq t \leq T'-1}\|\gamma_t\|_\infty \|x^* - x_t\|_1$, where $X_t := \langle \zeta_t, x^* - x_t\rangle$. By Lemma 2, we know that $X_t | X_1, \ldots, X_{t-1}$ is a centered sub-exponential random variable with parameters $\nu = \sqrt{n/2} \cdot \alpha \lesssim \sigma\|x^* - x_t\|_2/\delta\sqrt{n} \lesssim \sigma\|x^*\|_1/\delta\sqrt{n}$. 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 $1 - \mathcal{O}(d^{-1})$

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

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

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

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

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

$$\leq \frac{B^2\log d}{\eta}\frac{n}{T} + \frac{\sigma B}{\delta}\sqrt{\frac{\log d}{T}} + B\left(H\delta + \frac{\sigma s\log d}{\delta n}\right) + \eta\left(\frac{\sigma^2\log d}{\delta^2 n} + H^2\delta^2\right) \tag{21}$$

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 η and n, it is easy to verify that $\eta < 1/2H$, $n = \Omega(s^2\log d)$ and $n = \mathcal{O}(T)$. Subsequently,

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

Proof of Lemma 3

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

$$\begin{aligned}
\widetilde{g}_t(\delta) - g_t = \frac{\delta}{2}\mathbb{E}\left[(z^\top H_t z)z\right] + \underbrace{\frac{1}{n\delta}Z_t^\top\xi}_{:=\widetilde{\zeta}_t(\delta)} + \underbrace{\frac{\delta}{2n}\sum_{i=1}^n (z_i^\top H_t z_i)z_i - \mathbb{E}[(z^\top H_t z)z]}_{:=\widetilde{\beta}_t(\delta)}\\
+ \underbrace{\frac{\delta}{2n}\sum_{i=1}^n (z_i^\top(H_t(\delta z_i) - H_t)z_i)z_i + \left[(\widehat{\Sigma}-I)(\widehat{\theta}_t - \theta_{0t})\right]_{1:d}}_{:=\widetilde{\gamma}_t(\delta)},
\end{aligned}$$

where $\widehat{\Sigma}$, $\widehat{\theta}_t$ and θ_{0t} are similarly defined as in the proof of Lemma 2. The sub-exponentiality of $\langle\widetilde{\zeta}_t(\delta), a\rangle$ for any $a\in\mathbb{R}^d$ is established in Lemma 2. We next consider $\widetilde{\beta}_t(\delta)$. For any $a\in\mathbb{R}^d$ consider $\langle\widetilde{\beta}_t(\delta), a\rangle = \frac{\delta}{2n}\sum_{i=1}^n X_i(a)$ where $X_i(a) = (z_i^\top H_t z_i)(z_i^\top a) - \mathbb{E}[(z_i^\top H_t z_i)(z_i^\top a)]$ are centered i.i.d. random variables conditioned on H_t and x_t. In addition, $|X_i(a)| \leq 2\|H_t\|_1\|z_i\|_\infty^2 \cdot \|a\|_1\|z_i\|_\infty \lesssim H\|a\|_1$ almost surely. Therefore, $X_i(a)$ is a sub-Gaussian random variable with parameter $\nu = H\|a\|_1$, and hence $\langle\widetilde{\beta}_t(\delta), a\rangle$ is a sub-Gaussian random variable with parameter $\nu = \delta H\|a\|_1/\sqrt{n}$. Finally, for the deterministic term $\widetilde{\gamma}_t(\delta)$, we have that

$$\begin{aligned}
\|\widetilde{\gamma}_t(\delta)\|_\infty &\leq \frac{\delta}{2}\sup_{z\in\{\pm1\}^d}\|H_t(\delta z) - H_t\|_1\|z\|_\infty^2 + \|(\widehat{\Sigma}-I)(\widehat{\theta}_t - \theta_{0t})\|_\infty\\
&\leq \frac{\delta}{2}\sup_{z\in\{\pm1\}^d} L\cdot\|\delta z\|_\infty\|z\|_\infty^2 + \|\widehat{\Sigma}-I\|_{\max}\|\widehat{\theta}_t - \theta_{0t}\|_\infty\\
&\lesssim L\delta^2 + \sqrt{\frac{\log d}{n}}\left(\frac{\sigma s}{\delta}\sqrt{\frac{\log d}{n}} + s\delta H\right)\\
&\lesssim L\delta^2 + \frac{\sigma s\log d}{n\delta} + s\delta H\sqrt{\frac{\log d}{n}}.
\end{aligned}$$

Proof of Theorem 3

Because f is convex, $R_{\mathcal{A}}^{\mathsf{S}}(T) = f(x_{T+1}) - f^* \leq \frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*$. 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) $\|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_{\psi^*}^2$ and $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle\widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle$.

For the first term, using sub-exponentiality of $\widetilde{\zeta}_t$ and sub-gaussianity of $\widetilde{\beta}_t$, we have with probability $1-\mathcal{O}(d^{-1})$ uniformly over all $t \in \{0, \ldots, T'-1\}$,

$$\begin{aligned}
\|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_\infty &\le \|\widetilde{\zeta}_t\|_\infty + \|\widetilde{\beta}_t\|_\infty + \|\widetilde{\gamma}_t\|_\infty \\
&\lesssim \frac{\sigma}{\delta}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right) + \delta H\sqrt{\frac{\log d}{n}} + L\delta^2 + H\delta\sqrt{\frac{s^2\log d}{n}} + \frac{\sigma s \log d}{\delta n} \\
&\lesssim \left(\frac{\sigma}{\delta} + s\delta H\right)\sqrt{\frac{\log d}{n}} + L\delta^2,
\end{aligned}$$

where the last inequality holds because $n = \Omega(s^2 \log d)$. Subsequently, with probability $1-\mathcal{O}(d^{-1})$

$$\sup_{0\le t\le T'-1} \|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_{\psi^*}^2 \lesssim \left(\frac{\sigma^2}{\delta^2} + s^2\delta^2H^2\right)\frac{\log d}{n} + L^2\delta^4. \tag{22}$$

For the other term $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle$, again using concentration inequalities of sub-exponential/sub-Gaussian martingales and noting that $\|x^*-x_t\|_2 \le \|x^*-x_t\|_1 \le 2B$, we have

$$\begin{aligned}
\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle &= \frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{\zeta}_t + \widetilde{\beta}_t + \widetilde{\gamma}_t, x^* - x_t\rangle \\
&\lesssim \left(\frac{\sigma}{\delta} + s\delta H\right)B\sqrt{\frac{\log d}{T}} + B\left(L\delta^2 + \frac{\sigma s\log d}{\delta n} + s\delta H\sqrt{\frac{\log d}{n}}\right).
\end{aligned} \tag{23}$$

Subsequently, combining Eqs. (22,23) with Eq. (17) we have

$$\begin{aligned}
\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) &\lesssim \frac{B^2\log d}{\eta}\frac{n}{T} + \left(\frac{\sigma}{\delta} + s\delta H\right)B\sqrt{\frac{\log d}{T}} + (B+\eta)\left(L\delta^2 + \frac{\sigma s\log d}{\delta n} + s\delta H\sqrt{\frac{\log d}{n}}\right) \\
&\quad + \eta\left(\frac{\sigma^2}{\delta^2} + s^2\delta^2H^2\right)\frac{\log d}{n} + \eta L^2\delta^4.
\end{aligned} \tag{24}$$

We are now ready to prove Theorem 3. It is easy to verify that with the condition imposed on T and the selection of η and n, it holds that $\eta < 1/2H$, $n = \Omega(s^2\log d)$ and $n \le T/10$. Subsequently,

$$\begin{aligned}
&\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) \\
&\lesssim Bn^{1/3}\sqrt{\frac{\log d}{T}} + \left[\sigma\left(\frac{n}{s\log d}\right)^{1/3} + \widetilde{\mathcal{O}}(n^{-1/3})\right]B\sqrt{\frac{\log d}{T}} + \left(B + \widetilde{\mathcal{O}}\left(\frac{n^{2/3}}{\sqrt{T}}\right)\right)\left[(L+\sigma)\left(\frac{s\log d}{n}\right)^{2/3} + \widetilde{\mathcal{O}}(n^{-5/6})\right] \\
&+ Bn^{2/3}\sqrt{\frac{\log d}{T}}\left(\sigma^2\left(\frac{n}{s\log d}\right)^{2/3} + \widetilde{\mathcal{O}}(n^{-2/3})\right)\frac{\log d}{n} + Bn^{2/3}\sqrt{\frac{\log d}{T}}L^2\left(\frac{s\log d}{n}\right)^{4/3} \\
&\lesssim Bn^{1/3}\sqrt{\frac{\log d}{T}} + \sigma B\left(\frac{n}{s\log d}\right)^{1/3}\sqrt{\frac{\log d}{T}} + B(L+\sigma)\left(\frac{s\log d}{n}\right)^{2/3} + \sigma^2 B\left(\frac{n}{s^2\log^2 d}\right)^{1/3}\sqrt{\frac{\log d}{T}} + \widetilde{\mathcal{O}}(T^{-5/12}) \\
&\lesssim \left(B\sqrt{\log d} + \frac{\sigma B\sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B\sqrt{\log d}}{s^{2/3}}\right)\left[\frac{(1+L)s^{2/3}}{T}\right]^{1/3} + \frac{B(L+\sigma)}{(1+L)^{2/3}}\left(\frac{s^{2/3}\log d}{T}\right)^{1/3} + \widetilde{\mathcal{O}}(T^{-5/12}) \\
&\lesssim \left(B\sqrt{\log d} + \frac{\sigma B\sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B\sqrt{\log d}}{s^{2/3}}\right)\left[\frac{(1+L)s^{2/3}}{T}\right]^{1/3} + B\sigma\sqrt{\log d}\left(\frac{(1+L)s^{2/3}}{T}\right)^{1/3} + \widetilde{\mathcal{O}}(T^{-5/12}) \\
&\lesssim (1+\sigma+\sigma^2/s^{2/3})B\sqrt{\log d}\left(\frac{(1+L)s^{2/3}}{T}\right)^{1/3} + \widetilde{\mathcal{O}}(T^{-5/12}).
\end{aligned}$$

Additional tail inequalities

Lemma 6. *Suppose X and Y are centered sub-Gaussian random variables with parameters ν_1^2 and ν_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 ν_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,

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

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

Lemma 7 (Bernstein's inequality)**.** *Suppose X is a sub-exponential random variable with parameters ν and α.*

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

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 $\mathbb{E}[|X_j|^k|X_1, \ldots, X_{j-1}] \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for all integers $k \geq 3$. Then for all $t > 0$,*

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