A. Proofs for Realizable Setting

Proof of Lemma 3. Let $\Delta := \widehat{w} - w^*$ be the difference between the true answer and solution to the optimization problem. Let S to be the support of w^* and let $S^c = [d] \setminus S$ be the complements of S. Consider the permutation $i_1, \ldots, i_{d-k}$ of S^c for which $|\Delta(i_j)| \geq |\Delta(i_{j+1})|$ for all j. That is, the permutation dictated by the magnitude of the entries of Δ outside of S. We split S^c into subsets of size k according to this permutation: Define S_j, for $j \geq 1$ as $\{i_{(j-1)k+1}, \ldots, i_{jk}\}$. For convenience we also denote by S_{01} the set $S \cup S_1$.

Now, consider the matrix $X_{S_{01}} \in \mathbb{R}^{t \times |S_{01}|}$ whose columns are those of X with indices S_{01}. The Restricted Isometry Property of X dictates that for any vector $c \in \mathbb{R}^{S_{01}}$,

$$(1-\epsilon)\left\|c\right\|_2 \leq \frac{1}{\sqrt{n}}\left\|X_{S_{01}}c\right\|_2 \leq (1+\epsilon)\left\|c\right\|_2.$$

Let $V \subseteq \mathbb{R}^t$ be the subspace of dimension $|S_{01}|$ that is the image of the linear operator $X_{S_{01}}$, and let $P_V \in \mathbb{R}^{t \times t}$ be the projection matrix onto that subspace. We have, for any vector $z \in \mathbb{R}^t$ that

$$(1-\epsilon)\left\|P_V z\right\| \leq \frac{1}{\sqrt{n}}\left\|X_{S_{01}}^T z\right\| \leq (1+\epsilon)\left\|P_V z\right\|$$

We apply this to $z = X\Delta$ and conclude that

$$\left\|P_V X\Delta\right\| \leq \frac{1}{\sqrt{t}(1-\epsilon)}\left\|X_{S_{01}}^T X\Delta\right\| \tag{10}$$

We continue to lower bound the quantity of $\|P_V X\Delta\|$. We decompose $P_V X\Delta$ as

$$P_V X\Delta = P_V X\Delta(S_{01}) + \sum_{j\geq 2} P_V X\Delta(S_j) \tag{11}$$

Now, according to the definition of V we that there exist vectors $\{c_j\}_{j\geq 2}$ in $\mathbb{R}^{|S_{01}|}$ for which

$$P_V X\Delta(S_j) = X_{S_{01}}c_j$$

We now invoke Lemma 1.1 from (Candes & Tao, 2005) stating that for any S', S'' with $|S'| + |S''| \leq 3k$ it holds that

$$\forall c, c' \quad \frac{1}{n}\left\langle X_{S'}c, X_{S''}c'\right\rangle \leq (2\epsilon - \epsilon^2)\left\|c\right\|_2\left\|c'\right\|_2$$

We apply this for S_{01}, S_j, $j \geq 2$ and conclude that

$$\left\|P_V X\Delta(S_j)\right\|_2^2 = \left\langle P_V X\Delta(S_j), X\Delta(S_j)\right\rangle \leq 2\epsilon t\left\|c_j\right\|_2 \cdot \left\|\Delta(S_j)\right\| \leq \frac{2\epsilon\sqrt{t}}{1-\epsilon}\left\|P_V X\Delta(S_j)\right\|_2 \cdot \left\|\Delta(S_j)\right\|_2.$$

Dividing through by $\|P_V X\Delta(S_j)\|_2$, we get

$$\left\|P_V X\Delta(S_j)\right\| \leq \frac{2\epsilon\sqrt{t}}{1-\epsilon}\left\|\Delta(S_j)\right\|. \tag{12}$$

Let us now bound the sum $\|\Delta(S_j)\|$. By the definition of S_j we know that any element $i \in S_j$ has the property $\Delta(i) \leq (1/k)\left\|\Delta(S_{j-1})\right\|_1$. Hence

$$\sum_{j\geq 2}\left\|\Delta(S_j)\right\| \leq (1/\sqrt{k})\sum_{j\geq 1}\left\|\Delta(S_j)\right\|_1 = (1/\sqrt{k})\left\|\Delta(S^c)\right\|_1$$

We now combine this inequality with Equations (10), (11) and (12)

$$\begin{aligned}
\frac{1}{t}\left\|X_{S_{01}}^T X\Delta\right\| &\geq \frac{1-\epsilon}{\sqrt{t}}\left\|P_V X\Delta\right\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\left\|P_V X\Delta(S_{01})\right\| - \frac{1-\epsilon}{\sqrt{n}}\sum_{j\geq 2}\left\|P_V X\Delta(S_j)\right\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\left\|X\Delta(S_{01})\right\| - 2\epsilon\sum_{j\geq 2}\left\|\Delta(S_j)\right\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\left\|X\Delta(S_{01})\right\| - \frac{2\epsilon}{\sqrt{k}}\left\|\Delta(S^c)\right\|_1
\end{aligned}$$

The third inequality holds since $X\Delta(S_{01}) \in V$ hence $P_V X\Delta(S_{01)} = X\Delta(S_{01})$. We continue to bound the expression by claiming that $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$. This holds since in S^c, $\widehat{w}_{S^c} = \Delta(S^c)$ hence

$$\|w^*\|_1 = \|\widehat{w} - \Delta(S^c) - \Delta(S)\|_1 \leq \|\widehat{w}\|_1 + (\|\Delta(S)\|_1 - \|\Delta(S^c)\|_1)$$

Now, the optimality of $\widehat{w}$ implies $\|\widehat{w}\|_1 \leq \|w^*\|_1$, hence indeed $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$.

$$\|\Delta(S^c)\|_1 \leq \|\Delta(S)\|_1 \leq \sqrt{k}\,\|\Delta(S)\|_2 \leq \|\Delta(S_{01})\|_2 \leq \frac{\sqrt{k}}{(1-\epsilon)\sqrt{t}}\,\|X\Delta(S_{01})\|$$

We continue the chain of inequalities

$$\begin{aligned}\frac{1}{t}\left\|X_{S_{01}}^T X\Delta\right\| &\geq \frac{1-\epsilon}{\sqrt{n}}\,\|X\Delta(S_{01})\| - \frac{2\epsilon}{\sqrt{k}}\,\|\Delta(S^c)\|_1 \\ &\geq \|X\Delta(S_{01})\|\left(\frac{1-\epsilon}{\sqrt{n}} - \frac{2\epsilon}{\sqrt{k}}\cdot\frac{\sqrt{k}}{(1-\epsilon)\sqrt{n}}\right) \\ &= \frac{(1-\epsilon)^2 - 2\epsilon}{(1-\epsilon)\sqrt{t}}\,\|X\Delta(S_{01})\|\end{aligned}$$

Rearranging we conclude that

$$\begin{aligned}\|\Delta(S_{01})\| &\leq \frac{1}{(1-\epsilon)\sqrt{t}}\,\|X\Delta(S_{01})\| && \text{(RIP of } X\text{)} \\ &\leq \frac{1}{((1-\epsilon)^2 - 2\epsilon)t}\left\|X_{S_{01}}^T X\Delta\right\| \\ &\leq \frac{\sqrt{2k}}{(1-4\epsilon)t}\left\|X^T X\Delta\right\|_\infty && \text{(since for any } z\in\mathbb{R}^{2k},\ \|z\|_2 \leq \sqrt{2k}\,\|z\|_\infty\text{)} \\ &\leq C\sqrt{\frac{dk\log(d/\delta)}{tk_0}}\left(\sigma + \frac{d}{k_0}\,\|w^*\|_1\right) && \text{(Lemma 14 and } \epsilon < 1/5\text{)}\end{aligned}$$

for some constant C. We continue our bound on $\|\Delta\|$ by showing that $\|\Delta(S_{01}^c)\| \leq \|\Delta(S_{01})\|$

$$\|\Delta(S_{01}^c)\|_2^2 \overset{(i)}{\leq} \|\Delta(S^c)\|_1^2 \cdot \sum_{j\geq k+1}\frac{1}{j^2} \leq \frac{1}{k}\,\|\Delta(S^c)\|_1^2 \leq \frac{1}{k}\,\|\Delta(S)\|_1^2 \leq \|\Delta(S)\|_2^2\,.$$

Inequality (i) holds due to the following: Let α_i be the absolute value of the i'th largest (in absolute value) element of $\Delta(S^c)$. It obviously holds that $\alpha_i \leq \|\Delta(S^c)\|_1 / i$. Now, according to the definition of S_{01} we have that $\|\Delta(S_{01}^c)\|_2^2 = \sum_{j\geq k+1}\alpha_i^2$ and the inequality follows. Hence,

$$\|\Delta(S_{01}^c)\|_2 \leq \|\Delta(S)\|_2 \leq \|\Delta(S_{01})\|_2\,.$$

We conclude that

$$\|\Delta\|_2 \leq \sqrt{2}\,\|\Delta(S_{01})\|_2 \leq C\sqrt{\frac{dk\log(d/\delta)}{tk_0}}\left(\sigma + \frac{d}{k_0}\,\|w^*\|_1\right)$$

for some universal constant $C > 0$. Since $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$ and $|S| \leq k$ we get that

$$\|\Delta\|_1 \leq 2\,\|\Delta(S)\|_1 \leq 2\sqrt{k}\,\|\Delta(S)\|_2 \leq 2\sqrt{k}\,\|\Delta\|_2$$

and the claim follows. □

Proof of Lemma 4. Let S be the support of w^*. We can decompose the square of the left hand side as

$$\left\|\widehat{w}(\widetilde{S}) - w^*\right\|_2^2 = \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 + \sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2.$$

We upper bound the last sum on the right hand side as

$$\begin{aligned}
\sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2 &= \sum_{i \in S \setminus \widetilde{S}} [(\widehat{w}(i) - w^*(i)) + (\widehat{w}(i))]^2 \\
&\leq 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + (\widehat{w}(i))^2 \\
&\leq 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 ,
\end{aligned}$$

where first inequality follows from the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and the second inequality is due to the fact that $\widetilde{S}$ contains top k entries of $\widehat{w}$ in absolute value and $|S \setminus \widetilde{S}| = |\widetilde{S} \setminus S|$. Hence,

$$\begin{aligned}
\left\|\widehat{w}(\widetilde{S}) - w^*\right\|_2^2 &= \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 + \sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2 \\
&\leq \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&\leq 2 \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&= 2 \sum_{i \in S} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&\leq 3 \sum_{i=1}^{d} (\widehat{w}(i) - w^*(i))^2 \\
&= 3 \left\|\widehat{w} - w^*\right\|_2^2 .
\end{aligned}$$

Taking square root finishes the proof. □

Lemma 14. *There exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, the convex program (3) is feasible and its optimal solution $\widehat{w}$ satisfies*

$$\left\| \frac{1}{t} X_t^T X_t (\widehat{w} - w^*) \right\|_\infty \leq C \sqrt{\frac{d \log(d/\delta)}{t k_0}} \left(\sigma + \frac{d}{k_0} \left\| w^* \right\|_1 \right).$$

We note that the above lemma is beyond simple triangle inequality on the feasibility constraints, as the left hand side depends on actual design matrix X_t which we do not observe, instead of $\widehat{X}_t$.

Proof. To simplify notation, we drop subscript t. Namely, let $X = X_t$, $\widehat{X} = X_t$ and $\widehat{D} = \widehat{D}_t$, and also let $\eta = (\eta_1, \eta_2, \ldots, \eta_t)$ be the vector of noise variables.

First, we show that w^* satisfies the constraint of (3) with probability at least $1 - \delta$. We upper bound

$$\begin{aligned}
\left\| \frac{1}{t} \widehat{X}^T (Y - \widehat{X} w^*) + \frac{1}{t} \widehat{D} w^* \right\|_\infty &= \left\| \left[\frac{1}{t} \widehat{X}^T (X - \widehat{X}) + \frac{1}{t} \widehat{D} \right] w^* + \frac{1}{t} \widehat{X}^T \eta \right\|_\infty \\
&\leq \left\| \left[\frac{1}{t} \widehat{X}^T (X - \widehat{X}) + \frac{1}{t} \widehat{D} \right] w^* \right\|_\infty + \frac{1}{t} \left\| \widehat{X}^T \eta \right\|_\infty
\end{aligned}$$

We first bound the left summand. By Lemma 15, we have

$$\begin{aligned}\left\|\left[\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right]w^*\right\|_\infty &\le \|w^*\|_1\cdot\left\|\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right\|_\infty \\ &\le \|w^*\|_1\left(\left\|\frac{1}{t}X^T(\widehat{X}-X)\right\|_\infty+\left\|\frac{1}{t}(\widehat{X}-X)^T(\widehat{X}-X)-\frac{1}{t}\widehat{D}\right\|_\infty\right) \\ &\le \|w^*\|_1 C\cdot\sqrt{\frac{d^3\log(d/\delta)}{t{k_0}^3}}.\end{aligned}$$

For the right summand, since η is vector of i.i.d Gaussians with variance σ^2, with probability at least $1-\delta$,

$$\frac{1}{t}\left\|\widehat{X}^T\eta\right\|_\infty \le C\frac{\sigma}{t}\sqrt{\log(d/\delta)}\cdot\max_{i\in[d]}\left\|\widehat{X}_{(i)}\right\|_2$$

where $\widehat{X}_{(1)},\widehat{X}_{(2)},\ldots,\widehat{X}_{(d)}$ are the columns of $\widehat{X}$. Since the absolute value of the entries of $\widehat{X}$ is at most d/k_0, we have $\left\|\widehat{X}_{(i)}\right\|_2\le\sqrt{td/k_0}$ and thus

$$\frac{1}{t}\left\|\widehat{X}^T\eta\right\|_\infty \le C\sigma\sqrt{\frac{d\log(d/\delta)}{tk_0}}.$$

Combining the inequalities so far provides

$$\left\|\frac{1}{t}\widehat{X}^T(Y-\widehat{X}w^*)+\frac{1}{t}\widehat{D}w^*\right\|_\infty \le C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\left(\sigma+\frac{d}{k_0}\|w^*\|_1\right)$$

and hence conclude the constraint of the optimization problem (3) is satisfied (at least) by w^* and thus the optimization problem is feasible.

Now consider the vector $\Delta:=\widehat{w}-w^*$, we have

$$\begin{aligned}\left\|\frac{1}{t}X^TX\Delta\right\|_\infty &\le \left\|\frac{1}{t}(\widehat{X}^T\widehat{X}-\widehat{D})\Delta\right\|_\infty+\left\|\frac{1}{t}(\widehat{X}^T\widehat{X}-\widehat{D}-X^TX)\Delta\right\|_\infty \\ &\le \left\|\frac{1}{t}(\widehat{X}^T\widehat{X}-\widehat{D})\Delta\right\|_\infty+\left\|\frac{1}{t}(\widehat{X}-X)^TX\Delta\right\|_\infty \\ &\quad+\left\|\frac{1}{t}X^T(\widehat{X}-X)\Delta\right\|_\infty+\left\|\left(\frac{1}{t}(\widehat{X}-X)^T(\widehat{X}-X)-\frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty.\end{aligned}$$

According to Lemma 15 we have

$$\left\|\frac{1}{t}X^T(\widehat{X}-X)\Delta\right\|_\infty \le \left\|\frac{1}{t}X^T(\widehat{X}-X)\right\|_\infty\|\Delta\|_1 \le C\sqrt{\frac{d\log(d/\delta)}{tk_0}}(\|w^*\|_1+\|\widehat{w}\|_1)\le 2C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\cdot\|w^*\|_1$$

where the last inequality is by the optimality of $\widehat{w}$. The same argument provides an identical bound for $\left\|\frac{1}{t}(\widehat{X}-X)^TX\Delta\right\|_\infty$. The last summand can also be bounded by using Lemma 15 and the optimality of $\widehat{w}$.

$$\left\|\left(\frac{1}{t}(\widehat{X}-X)^T(\widehat{X}-X)-\frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty \le 2C\sqrt{\frac{d^3\log(d/\delta)}{t{k_0}^3}}\cdot\|w^*\|_1$$

Finally, according to the feasibility of $\widehat{w}$ and w^* we may bound the first summand

$$\left\|\left(\frac{1}{t}\widehat{X}^T\widehat{X}-\frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty \le 2C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\left(\sigma+\frac{d}{k_0}\|w^*\|_1\right),$$

and reach the final bound. □

Lemma 15. *For any $t \geq t_0$, with probability at least $1-\delta$, the following two inequalities hold*

$$\left\| \frac{1}{t}(\widehat{X}_t - X_t)^T(\widehat{X}_t - X_t) - \frac{1}{t}\widehat{D}_t \right\|_\infty \leq C\sqrt{\frac{d^3 \log(d/\delta)}{t{k_0}^3}}\,,$$
$$\left\| \frac{1}{t}X_t^T(\widehat{X}_t - X_t) \right\|_\infty \leq C\sqrt{\frac{d \log(d/\delta)}{tk_0}}\,,$$

where $\|\cdot\|_\infty$ denotes the maximum of the absolute values of the entries of a matrix.

Proof. Throughout we use that $|x_s(i)| \leq 1$ for all $i \in [d]$ and all $s \in [t]$, and (2) $(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}$ is unbiased with absolute value of at most $(d/k_0)^2$ and variance of at most $(d/k_0)^3$. For the first term, let's bound

$$\left[\frac{1}{t}(\widehat{X} - X)^T(\widehat{X} - X) - \frac{1}{t}\widehat{D}\right]_{ij} = \frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))(\widehat{x}_s(j) - x_s(j)) - \frac{1}{t}\widehat{D}_{ij}$$

For $i = j$, we have

$$\mathbf{E}\left[\left((\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right)^2\right] \leq \mathbf{E}\left[(\widehat{x}_s(i) - x_s(i))^4\right] \leq (d/k_0)^3$$

$$(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii} \leq (d/k_0)^2, \quad \mathbf{E}\left[(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right] = 0$$

Hence, by Bernstein's inequality, for any $v > 0$,

$$\Pr\left[\left|\frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right| > v\right] \leq 2\exp\left(-\frac{v^2 t}{(d/k_0)^3 + (d/k_0)^2 v/3}\right).$$

It follows that for any $\delta > 0$, with probability at least $1-\delta$ it holds for all $i \in [d]$ that,

$$\left|\frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right| \leq \mathcal{O}\left(\frac{\log(d/\delta)d^2}{t{k_0}^2} + \sqrt{\frac{\log(d/\delta)d^3}{t{k_0}^3}}\right).$$

Similarly we have $\frac{1}{t}(\widehat{D}_{ii} - D_{ii}) \leq \mathcal{O}\left(\frac{\log(d/\delta)d^2}{t{k_0}^2} + \sqrt{\frac{\log(d/\delta)d^3}{t{k_0}^3}}\right)$.

For $i \neq j$ we use an analogous argument, only now the variance term in Bernstein's inequality is $(d/k_0)^2$ rather than $(d/k_0)^3$, hence only reach a tighter bound.

For the second term, we again bound via Bernstein's inequality as

$$\left[\frac{1}{t}X^T(\widehat{X} - X)\right]_{ij} = \frac{1}{t}\sum_{s=1}^{t}x_s(i)(\widehat{x}_s(j) - x_s(j)) \leq \mathcal{O}\left(\sqrt{\frac{d\log(d/\delta)}{tk_0}} + \frac{d\log(d/\delta)}{tk_0}\right)$$

The claim now follows by noticing that for large enough t, the dominating terms are those that scale as $1/\sqrt{t}$. □

Proof of Theorem 2. By Lemma 3,

$$\|w_{t+1} - w^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d}{k_0}\frac{k\log(d/\delta)}{t}}(\sigma + \frac{d}{k_0}\|w^*\|_1)\right).$$

We have

$$\begin{aligned}\mathrm{Regret}_T(w^*) - \mathrm{Regret}_{t_0}(w^*) &= \sum_{t=t_0+1}^{T} (y_t - \langle x_t, w_t\rangle)^2 - (y_t - \langle x_t, w^*\rangle)^2 \\ &= \sum_{t=t_0+1}^{T} (\langle x_t, w^* - w_t\rangle + \eta_t)^2 - \eta_t^2 \\ &= \sum_{t=t_0+1}^{T} (\langle x_t, w^* - w_t\rangle + 2\eta_t) \langle x_t, w^* - w_t\rangle \\ &= \sum_{t=t_0+1}^{T} 2\eta_t \langle x_t, w^* - w_t\rangle + (\langle x_t, w^* - w_t\rangle)^2 \ ,\end{aligned}$$

where we used that $y_t = \langle x_t, w_t\rangle + \eta_t$. To bound the regret we require the upper bound, that occurs with probability of at least $1-\delta$,

$$\forall t \geq t_0 \qquad |\langle x_t, w^* - w_t\rangle| \overset{(i)}{\leq} \|x_t\|_\infty \sqrt{\|w_t - w^*\|_0} \cdot \|w_t - w^*\|_2 \overset{(ii)}{\leq} \mathcal{O}\left(k \cdot \sqrt{\frac{d}{k_0} \frac{\log(\log(T)d/\delta)}{t}} \left(\sigma + \frac{d}{k_0}\right)\right).$$

Inequality (i) holds since $\langle a, b\rangle \leq \|a(S)\|_2 \cdot \|b\|_2$ with S being the support of b and $\|a(S)\|_2 \leq \|a\|_\infty \sqrt{|S|}$. Inequality (ii) follows from Lemma 3 and Lemma 4, and a union bound over the $\lceil \log(T) \rceil$ many times the vector w_t is updated. Now, for the left summand of the regret bound we have by Martingale concentration inequality that w.p. $1-\delta$

$$\begin{aligned}\sum_{t=t_0+1}^{T} 2\eta_t \langle x_t, w_t - w^*\rangle &\leq \mathcal{O}\left(\sigma \sqrt{\log(1/\delta) \sum_{t=t_0+1}^{T} \langle x_t, w_t - w^*\rangle^2}\right) \\ &= \mathcal{O}\left(\sigma \sqrt{\log(1/\delta)\log(T) k^2 \cdot \frac{d \log(d\log(T)/\delta)}{k_0} \left(\sigma + \frac{d}{k_0}\right)^2}\right).\end{aligned}$$

The right summand is bounded as

$$\sum_{t=t_0+1}^{T} \langle x_t, w^* - w_t\rangle^2 = \mathcal{O}\left(k^2 \cdot \frac{d \log(d\log(T)/\delta)}{k_0} \left(\sigma + \frac{d}{k_0}\right)^2 \cdot \log(T)\right).$$

Clearly, the right summand dominates the left one.

It remains to bound the regret in first t_0 rounds. Since $w_t = 0$ for $t \leq t_0$, we have

$$\mathrm{Regret}_{t_0}(w^*) = \sum_{t=1}^{t_0} 2\eta_t \langle x_t, w^*\rangle + (\langle x_t, w^*\rangle)^2 \leq \mathcal{O}\left(\sigma\sqrt{t_0 \log(1/\delta)} + t_0\right).$$

Here, we used that $|\langle x_t, w^*\rangle| \leq 1$ since $\|x_t\|_\infty \leq 1$ and $\|w^*\|_1 \leq 1$. We also used that $\eta_t \langle x_t, w^*\rangle \sim N(0, \sigma^2 \langle x_t, w^*\rangle^2)$ and $\eta_1 \langle x_1, w^*\rangle, \eta_2 \langle x_2, w^*\rangle, \ldots, \eta_{t_0} \langle x_{t_0}, w^*\rangle$ are independent. Thus their sum is a Gaussian with variance at most $\sigma^2 t_0$.

Collecting all the terms along with Lemma 16, bounding the difference $\mathrm{Regret}_T - \mathrm{Regret}_T(w^*)$, gives

$$\mathrm{Regret}_T \leq \left(t_0 + \sqrt{t_0 \log(1/\delta)} + k^2 \cdot \frac{d \log(d\log(T)/\delta)}{k_0} \left(\sigma + \frac{d}{k_0}\right)^2 \cdot \log(T)\right) \tag{13}$$

□

Lemma 16. *In the realizable case, w.p. at least $1-\delta$ we have for any sequence of w_t that $\mathrm{Regret}_T - \mathrm{Regret}_T(w^*) = O(\sigma^2 k \log(d/\delta))$.*

Proof. It is an easy exercise to show that $\text{Regret}_T - \text{Regret}_T(w^*)$ is equal to the regret on an algorithm that always plays w^*. We thus continue to bound the regret of w^*.

Let $\Delta \in \mathbb{R}^d$ be the difference between w^* and $\tilde{w}$, the empirical optimal solution for the sparse regression problem. The loss associated with w^* is clearly $\|\eta\|^2$, where η is the noise term $y = Xw^* + \eta$. The loss associated with $\tilde{w}$ is

$$\|X(w^* + \Delta) - Xw^* - \eta\|^2 = \|\eta - X\Delta\|^2 = \|\eta - X_{\tilde{S}}\Delta\|^2$$

where $\tilde{S}$ is the support of Δ, having a cardinality of at most $2k$. The closed form solution for the least-squares problem dictates that

$$\|\eta - X_{\tilde{S}}\Delta\|^2 \geq \|\eta - X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2 = \|\eta\|^2 - \|X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2 \,.$$

Here, $A^{\dagger}$ is the pseudo inverse of a matrix A and X_S is the matrix obtained from the columns of X whose indices are in S. It follows that the regret of w^* is bounded by

$$\|X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2$$

for some subset $\tilde{S}$ of size at most $2k$. To bound this quantity we use a high probability bound for $\|X_S X_S^{\dagger}\eta\|^2$ for a fixed set S, and take a union bound over all possible sets of cardinality $2k$. For a fixed set S we have that $\|X_S X_S^{\dagger}\eta\|^2/\sigma^2$ is distributed according to the χ^2_{2k} distribution. The tail bounds of this distribution suggest that

$$\Pr\left[\|X_S X_S^{\dagger}\eta\|^2 > 2k\sigma^2 + 2\sigma^2\sqrt{2kx} + 2\sigma^2 x\right] \leq \exp(-x)$$

meaning that with probability at least $1 - \delta/d^{2k}$ we have

$$\|X_S X_S^{\dagger}\eta\|^2 < 2k\sigma^2 + 2\sigma^2\sqrt{2k \cdot 2k \cdot \log(d/\delta)} + 2\sigma^2 \cdot 2k \cdot \log(d/\delta) = O(\sigma^2 k \log(d/\delta))$$

Taking a union bound over all possible subsets of size $\leq 2k$ we get that w.p. at least $1 - \delta$ the regret of w^* is at most $O(\sigma^2 k \log(d/\delta))$. □

B. Proofs for Agnostic Setting

We begin with an auxiliary lemma for Lemma 10, informally proving that for any matrix $\bar{X}$ with BBRCNP (Definition 6) and vector y, the set function

$$g(S) = \inf_{w \in \mathbb{R}^S} \|y - \bar{X}w\|^2$$

is weakly supermodular. Its proof can be found in (Boutsidis et al., 2015), yet for completeness we provide it here as well.

Lemma 17. *[Lemma 5 in (Boutsidis et al., 2015)] Let $\bar{X}$ be a matrix whose columns have 2-norm at most 1 and y be a vector with $\|y\|_\infty \leq 1$ of dimension matching the number of rows in X. the set function*

$$g(S) = \inf_{w \in \mathbb{R}^S} \|y - Xw\|^2$$

is α-weakly supermodular for sparsity k for $\alpha = \max_{S:|S|\leq k} 1/\sigma_{\min}(X_S)^2$, where X_S is the submatrix of X obtained by choosing the columns indexed by S, and $\sigma_{\min}(A)$ is the smallest singular value of A.

Proof. Firstly, the well known closed form solution for the least-squares problem informs us that

$$\begin{aligned} g(S) &= \inf_{w \in \mathbb{R}^S} \|y - Xw\|^2, \\ &= y^T[I - (X_S^T)^{\dagger}X_S^T]y. \end{aligned}$$

We use the notation $A^{\dagger}$ for the pseudoinverse of a matrix A. That is, if the singular value decomposition of A is $A = \sum_i \sigma_i u_i v_i^T$ with $\sigma_i > 0$ then $A^{\dagger} = \sum_i \sigma_i^{-1} v_i u_i^T$.

Let us first estimate $g(S) - g(T)$, for sets $S \subset T$. For brevity, define H_S as the projection matrix $X_S X_S^{\dagger}$ projecting onto the column space of X_S. Denote by $Z_{T\setminus S}$ the matrix whose columns are those of $X_{T\setminus S}$ projected away from the span of X_S, and normalized. Namely, writing x_i as the i'th column of X, $\zeta_i = \|(I - H_S)x_i\|$, $z_i = (I - H_S)x_i/\zeta_i$, and $Z_{T\setminus S}$'s

columns are $\{z_i\}_{i\in T\setminus S}$. Notice that the columns of $Z_{T\setminus S}$ and X_S are orthogonal, hence according to the Pythagorean theorem it holds that

$$g(S) = \|y\|^2 - \|H_S y\|^2, \quad g(T) = \|y\|^2 - \|H_S y\|^2 - \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2$$

meaning that $g(S)-g(T) = \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2$. In particular, this implies that for any $j \notin S$ it holds that $g(S)-g(S\cup\{j\}) = (z_j^T y)^2$, since z_j is a unit vector. Let us now decompose $g(S) - g(T)$.

$$g(S) - g(T) = \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2 = \|(Z_{T\setminus S}^T)^{\dagger} Z_{T\setminus S}^T y\|^2 \le \|(Z_{T\setminus S}^T)^{\dagger}\|^2 \cdot \|Z_{T\setminus S}^T y\|^2$$

The norm used in the last inequality is the matrix operator norm. We now bound both factors of the product on the RHS separately. For the first factor, we claim that $\|(Z_{T\setminus S}^T)^{\dagger}\| = \|Z_{T\setminus S}^{\dagger}\| \le \|X_T^{\dagger}\|$. To see this, consider a vector $w \in \mathbb{R}^{|T\setminus S|}$, for convenience denote its entries by $\{w(i)\}_{i\in T\setminus S}$, and write $z_i = (x_i - \sum_{j\in S} \alpha_{ij} x_j)/\zeta_i$. We have

$$Z_{T\setminus S} w = \sum_{i\in T\setminus S} z_i w(i) = \sum_{i\in T\setminus S} x_i w(i)/\zeta_i - \sum_{j\in S} x_j \sum_{i\in T\setminus S} w(i)\alpha_{ij}/\zeta_i = X_T w'$$

for the vector $w' \in \mathbb{R}^{|T|}$ defined as $w'(i) = w(i)/\zeta_i$ for $i \in T \setminus S$ and $w'(j) = -\sum_{i\in T\setminus S} w(i)\alpha_{ij}/\zeta_i$ for $j \in S$. Since $\zeta_i \le \|x_i\| \le 1$ we must have $\|w'\| \ge \|w\|$. Consider now the unit vector w for which $\|Z_{T\setminus S} w\| = \|Z_{T\setminus S}^{\dagger}\|^{-1}$, that is, the unit norm singular vector corresponding to the smallest non-zero singular value of $Z_{T\setminus S}$. For this w, and its corresponding vector w', we have

$$\|Z_{T\setminus S}^{\dagger}\|^{-1} = \|Z_{T\setminus S} w\| = \|X_T w'\| \ge \sigma_{\min}(X_T)\|w'\| \ge \sigma_{\min}(X_T)\|w\| = \sigma_{\min}(X_T).$$

It follows that

$$\|(Z_{T\setminus S}^T)^{\dagger}\|^2 = \|Z_{T\setminus S}^{\dagger}\|^2 \le 1/\sigma_{\min}(X_T)^2$$

We continue to bound the right factor of product.

$$\|Z_{T\setminus S}^T y\|^2 = \sum_{i\in T\setminus S} (z_i^T y)^2 = \sum_{i\in T\setminus S} g(S) - g(S\cup\{i\}).$$

By combining the inequalities we obtained the required result:

$$g(S) - g(T) \le \left(1/\sigma_{\min}(X_T)^2\right) \sum_{i\in T\setminus S} g(S) - g(S\cup\{i\}).$$

□

Proof of Lemma 10. We would like to apply Lemma 17 on the design matrix X. The only catch is that the columns of X may not be bounded by 1 in norm. To remedy this, let j be the index of the column with the maximum norm and consider the matrix $\bar{X} = \frac{1}{\|X_j\|} X$ instead (here, X_j is the j-th column of X; note that $X_j = Xe_j$ for the j-th standard basis vector e_j). Now, for any subset S of coordinates,

$$\inf_{w\in\mathbb{R}^S} \|y - \bar{X} w\|^2 = \inf_{w\in\mathbb{R}^S} \|y - Xw\|^2.$$

Thus, we conclude that the set function of interest, $g(S) = \inf_{w\in\mathbb{R}^S} \|y - Xw\|^2$, is α-weakly supermodular for sparsity k for $\alpha = \max_{S:|S|\le k} \|\bar{X}_S^{\dagger}\|_2^2$. For any subset of coordinates S of size at most k, let w be a unit norm right singular vector of $\bar{X}_S$ corresponding to the smallest singular value, so that $\|\bar{X}_S^{\dagger}\|_2 = \frac{1}{\|\bar{X}_S w\|}$. But $\frac{1}{\|\bar{X}_S w\|} = \frac{\|Xe_j\|}{\|Xw'\|}$, where w' is the vector w extended to all coordinates by padding with zeros.

Since the restricted condition number of X for sparsity k is bounded by κ we conclude that $\frac{\|Xe_j\|}{\|Xw'\|} \le \kappa$. Since this bound holds for any subset S of size at most k, we conclude that $\alpha \le \kappa^2$. □

Proof of Lemma 11. By the α-weak supermodularity of g, we have

$$\begin{aligned} g(\emptyset) - g(V) &\le \alpha \cdot \sum_{j \in V} [g(\emptyset) - g(\{j\})] \\ &\le \alpha |V| \cdot [(g(\emptyset) - g(V)) - (g(\{j^*\}) - g(V))]. \end{aligned}$$

Rearranging, we get the claimed bounds. □

The following lemma gives a useful property of weakly supermodular functions.

Lemma 18. *Let $g(\cdot)$ be a (k, α)-weakly supermodular set function and U be a subset with $|U| < k$. Then $g'(S) := g(U \cup S)$ is $(k - |U|, \alpha)$-weakly supermodular.*

Proof. For any two subsets $S \subseteq T$ with $|T| \le k - |U|$, we have

$$\begin{aligned} g'(S) - g'(T) = g(U \cup S) - g(U \cup T) &\le \alpha \sum_{j \in (T \cup U) \setminus (S \cup U)} [g(U \cup S) - g(U \cup S \cup \{j\})] \\ &\le \alpha \sum_{j \in T \setminus S} [g(U \cup S) - g(U \cup S \cup \{j\})] = \alpha \sum_{j \in T \setminus S} [g'(S) - g'(S \cup \{j\})]. \end{aligned}$$

□

Proof of Lemma 12. For $i \in \{0, 1, \ldots, k_1\}$, define the set function $g_b^{(i)}$ as $g_b^{(i)}(S) = g_b(S \cup V_b^{(i)})$.

First, we analyze the performance of the BEXP algorithms. Fix any $i \in [k_1]$ and consider $\mathsf{BEXP}^{(i)}$. Conceptually, for any $j \in [d]$, the loss of expert j at the end of mini-batch b is $g_b(V_b^{(i-1)} \cup j)$ (note that this loss is only evaluated for $j \in U_b^{(i)}$ in the algorithm). To bound the regret, we need to bound the magnitude of the losses. Note that for any subset S, we have $0 \le g_b(S) \le \frac{1}{B} \sum_{t \in \mathcal{T}_b} y_t^2 \le 1$. Thus, the regret guarantee of BEXP (Theorem 8) implies that for any $i \in [k_1]$ and any $j \in [d]$, we have

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b(V_b^{(i-1)} \cup \{j_b^{(i)}\})\right] \le \sum_{b=1}^{T/B} g_b(V_b^{(i-1)} \cup \{j\}) + 2\sqrt{\tfrac{dk_1 \log(d)T}{k_0 B}}.$$

The expectation above is conditioned on the randomness in $V_b^{(i-1)}$, for $b \in [T/B]$. Rewriting the above inequality using the $g^{(i-1)}$ and $g^{(i)}$ functions, and using the fact that $V_b^{(i-1)} \cup \{j_b^{(i)}\} = V_b^{(i)}$, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(i)}(\emptyset)\right] \le \sum_{b=1}^{T/B} g_b^{(i-1)}(\{j\}) + 2\sqrt{\tfrac{dk_1 \log(d)T}{k_0 B}}. \tag{14}$$

Next, since we assumed that the sequence of feature vectors satisfies BBRCNP with parameters $(\epsilon, k_1 + k)$, Lemma 10 implies that the set function g_b defined in (6) is $(k_1 + k, \kappa^2)$-weakly supermodular for $\kappa = \frac{1+\epsilon}{1-\epsilon}$. By Lemma 18, the set function $g_b^{(i)}$ is (k, κ^2)-weakly supermodular (since $|V_b^{(i)}| \le k_1$).

It is easy to check that the sum of weakly supermodular functions is also weakly supermodular (with the same parameters), and hence $\sum_{b=1}^{T/B} g_b^{(i-1)}$ is also (k, κ^2)-weakly supermodular. Hence, by Lemma 11, if $j^* = \arg\min_j \sum_{b=1}^{T/B} g_b^{(i-1)}(\{j\})$, we have, for any subset V of size at most k,

$$\sum_{b=1}^{T/B} g_b^{(i-1)}(\{j^*\}) - g_b^{(i-1)}(V) \le (1 - \tfrac{1}{\kappa^2 |V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b^{(i-1)}(V)].$$

Since $g_b(V) \ge g_b(V \cup V_b^{(i-1)}) = g_b^{(i-1)}(V)$, the above inequality implies that

$$\sum_{b=1}^{T/B} g_b^{(i-1)}(\{j^*\}) - g_b(V) \le (1 - \tfrac{1}{\kappa^2 |V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b(V)].$$

Combining this bound with (14) for $j = j^*$, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(i)}(\emptyset) - g_b(V)\right] \leq (1 - \tfrac{1}{\kappa^2|V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b(V)] + 2\sqrt{\tfrac{dk_1\log(d)T}{k_0 B}}.$$

Applying this bound recursively for $i \in [k_1]$ and simplifying, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(k_1)}(\emptyset) - g_b(V)\right] \leq (1 - \tfrac{1}{\kappa^2|V|})^{k_1}[\sum_{b=1}^{T/B} g_b^{(0)}(\emptyset) - g_b(V)] + 2\kappa^2|V|\sqrt{\tfrac{dk_1\log(d)T}{k_0 B}}.$$

Using the definitions of $g_b^{(k_1)}$ and $g_b^{(0)}$, and the fact that $|V| \leq k$, we get the claimed bound. □