

Adaptive Feature Selection: Computationally Efficient Online Sparse Linear Regression under RIP

Satyen Kale*
Google Research, New York
satyenkale@google.com

Zohar Karnin*
Amazon, New York
zkarnin@gmail.com

Tengyuan Liang*
University of Chicago
Booth School of Business
Tengyuan.Liang@chicagobooth.edu

Dávid Pál
Yahoo Research, New York
dpal@yahoo-inc.com

Abstract

Online sparse linear regression is an online problem where an algorithm repeatedly chooses a subset of coordinates to observe in an adversarially chosen feature vector, makes a real-valued prediction, receives the true label, and incurs the squared loss. The goal is to design an online learning algorithm with sublinear regret to the best sparse linear predictor in hindsight. Without any assumptions, this problem is known to be computationally intractable. In this paper, we make the assumption that data matrix satisfies restricted isometry property, and show that this assumption leads to computationally efficient algorithms with sublinear regret for two variants of the problem. In the first variant, the true label is generated according to a sparse linear model with additive Gaussian noise. In the second, the true label is chosen adversarially.

1 Introduction

In modern real-world sequential prediction problems, samples are typically high dimensional, and construction of the features may itself be a computationally intensive task. Therefore in sequential prediction, due to the computation and resource constraints, it is preferable to design algorithms that compute only a limited number of features for each new data example. One example of this situation, from [Cesa-Bianchi et al., 2011], is medical diagnosis of a disease, in which each feature is the result of a medical test on the patient. Since it is undesirable to subject a patient to a battery of medical tests, we would like to adaptively design diagnostic procedures that rely on only a few, highly informative tests.

Online sparse linear regression (OSLR) is a sequential prediction problem in which an algorithm is allowed to see only a small subset of coordinates of each feature vector. The problem is parameterized by 3 positive integers: d , the dimension of the feature vectors, k , the sparsity of the linear regressors we compare the algorithm's performance to, and k_0 , a budget on the number of features that can be queried in each round by the algorithm. Generally we have $k \ll d$ and $k_0 \geq k$ but not significantly larger (our algorithms need¹ $k_0 = \tilde{O}(k)$).

In the OSLR problem, the algorithm makes predictions over a sequence of T rounds. In each round t , nature chooses a feature vector $x_t \in \mathbb{R}^d$, the algorithm chooses a subset of $\{1, 2, \dots, d\}$ of size at most k' and observes the corresponding coordinates of the feature vector. It then makes a prediction $\hat{y}_t \in \mathbb{R}$ based on the observed features, observes the true label y_t , and suffers loss $(y_t - \hat{y}_t)^2$. The goal of the learner is

*This work was done while the author was at Yahoo Research, New York.

¹In this paper, we use the $\tilde{O}(\cdot)$ notation to suppress factors that are polylogarithmic in the natural parameters of the problem.

to make the cumulative loss comparable to that of the best k -sparse linear predictor w in hindsight. The performance of the online learner is measured by the *regret*, which is defined as the difference between the two losses:

$$\text{Regret}_T = \sum_{t=1}^T (y_t - \hat{y}_t)^2 - \min_{w: \|w\|_0 \leq k} \sum_{t=1}^T (y_t - \langle x_t, w \rangle)^2 .$$

The goal is to construct algorithms that enjoy regret that is sub-linear in T , the total number of rounds. A sub-linear regret implies that in the asymptotic sense, the average per-round loss of the algorithm approaches the average per-round loss of the best k -sparse linear predictor.

Sparse regression is in general a computationally hard problem. In particular, given k, x_1, x_2, \dots, x_T and y_1, y_2, \dots, y_T as inputs, the offline problem of finding a k -sparse w that minimizes the error $\sum_{t=1}^T (y_t - \langle x_t, w \rangle)^2$ does not admit a polynomial time algorithm under standard complexity assumptions Foster et al. [2015]. This hardness persists even under the assumption that there exists a k -sparse w^* such that $y_t = \langle x_t, w^* \rangle$ for all t . Furthermore, the computational hardness is present even when the solution is required to be only $\tilde{\mathcal{O}}(k)$ -sparse solution and has to minimize the error only approximately; see Foster et al. [2015] for details. The hardness result was extended to online sparse regression by Foster et al. [2016]. They showed that for all $\delta > 0$ there exists no polynomial-time algorithm with regret $\mathcal{O}(T^{1-\delta})$ unless $NP \subseteq BPP$.

Foster et al. [2016] posed the open question of what additional assumptions can be made on the data to make the problem tractable. In this paper, we answer this open question by providing efficient algorithms with sublinear regret under the assumption that the matrix of feature vectors satisfies the *restricted isometry property* (RIP) [Candes and Tao, 2005]. It has been shown that if RIP holds and there exists a sparse linear predictor w^* such that $y_t = \langle x_t, w^* \rangle + \eta_t$ where η_t is independent noise, the offline sparse linear regression problem admits computationally efficient algorithms, e.g., Candes and Tao [2007]. RIP and related Restricted Eigenvalue Condition [Bickel et al., 2009] have been widely used as a standard assumption for theoretical analysis in the compressive sensing and sparse regression literature, in the offline case. In the online setting, it is natural to ask whether sparse regression avoids the computational difficulty under an appropriate form of the RIP condition. In this paper, we answer this question in a positive way, both in the realizable setting and in the agnostic setting. As a by-product, we resolve the adaptive feature selection problem as the efficient algorithms we propose in this paper adaptively choose a different “sparse” subset of features to query at each round. This is closely related to attribute-efficient learning (see discussion in Section 1.2) and online model selection.

1.1 Summary of Results

We design polynomial-time algorithms for online sparse linear regression for two models for the sequence $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)$. The first model is called the *realizable* and the second is called *agnostic*. In both models, we assume that, after proper normalization, for all large enough t , the matrix X_t formed from the first t feature vectors x_1, x_2, \dots, x_t satisfies the restricted isometry property. The two models differ in the assumptions on y_t . The realizable model assumes that $y_t = \langle x_t, w^* \rangle + \eta_t$ where w^* is k -sparse and η_t is an independent noise. In the agnostic model, y_t can be arbitrary, and therefore, the regret bounds we obtain are worse than in the realizable setting. The models and corresponding algorithms are presented in Sections 2 and 3 respectively. Interestingly enough, the algorithms and their corresponding analyses are completely different in the realizable and agnostic case.

Our algorithms allow for somewhat more flexibility than the problem definition: they are designed to work with a budget k_0 on the number of features that can be queried that may be larger than the sparsity parameter k of the comparator. The regret bounds we derive improve with increasing values of k_0 . In the case when $k_0 \approx k$, the dependence on d in the regret bounds is polynomial, as can be expected in limited feedback settings (this is analogous to polynomial dependence on d in *bandit* settings). In the extreme case when $k_0 = d$, i.e. we have access to *all* the features, the dependence on the dimension d in the regret bounds we prove is only *logarithmic*. The interpretation is that if we have full access to the features, but the goal is to compete with just k sparse linear regressors, then the number of data points that need to be seen to achieve good predictive accuracy has only logarithmic dependence on d . This is analogous to the (offline)

compressed sensing setting where the sample complexity bounds, under RIP, only depend logarithmically on d .

A major building block in the solution for the realizable setting (Section 2) consists of identifying the best k -sparse linear predictor for the past data at any round in the prediction problem. This is done by solving a sparse regression problem on the observed data. The solution of this problem cannot be obtained by a simple application of say, the Dantzig selector [Candes and Tao, 2007] since we do not observe the data matrix X , but rather a subsample of its entries. Our algorithm is a variant of the Dantzig selector that incorporates random sampling into the optimization, and computes a near-optimal solution by solving a linear program. The resulting algorithm has a regret bound of $\tilde{O}(\log T)$. This bound has optimal dependence on T , since even in the full information setting where all features are observed there is a lower bound of $\Omega(\log T)$ [Hazan and Kale, 2014].

The algorithm for the agnostic setting relies on the theory of submodular optimization. The analysis in [Boutsidis et al., 2015] shows that the RIP assumption implies that the set function defined as the minimum loss achievable by a linear regressor restricted to the set in question satisfies a property called *weak supermodularity*. Weak supermodularity is a relaxation of standard supermodularity that is still strong enough to show performance bounds for the standard greedy feature selection algorithm for solving the sparse regression problem. We then employ a technique developed by Streeter and Golovin [2008] to construct an online learning algorithm that mimics the greedy feature selection algorithm. The resulting algorithm has a regret bound of $\tilde{O}(T^{2/3})$. It is unclear if this bound has the optimal dependence on T : it is easy to prove a lower bound of $\Omega(\sqrt{T})$ on the regret using standard arguments for the multiarmed bandit problem.

1.2 Related work

A related setting is attribute-efficient learning [Cesa-Bianchi et al., 2011, Hazan and Koren, 2012, Kukliansky and Shamir, 2015]. This is a batch learning problem in which the examples are generated i.i.d., and the goal is to simply output a linear regressor using only a limited number of features per example with bounded excess risk compared to the optimal linear regressor, when given *full access* to the features at test time. Since the goal is not prediction but simply computing the optimal linear regressor, efficient algorithms exist and have been developed by the aforementioned papers.

Without any assumptions, only inefficient algorithms for the online sparse linear regression problem are known Zolghadr et al. [2013], Foster et al. [2016]. Kale [2014] posed the open question of whether it is possible to design an efficient algorithm for the problem with a sublinear regret bound. This question was answered in the negative by Foster et al. [2016], who showed that efficiency can only be obtained under additional assumptions on the data. This paper shows that the RIP assumption yields tractability in the online setting just as it does in the batch setting.

In the realizable setting, the linear program at the heart of the algorithm is motivated from Dantzig selection Candes and Tao [2007] and error-in-variable regression Rosenbaum and Tsybakov [2010], Belloni et al. [2016]. The problem of finding the best sparse linear predictor when only a sample of the entries in the data matrix is available is also discussed by Belloni et al. [2016] (see also the references therein). In fact, these papers solve a more general problem where we observe a matrix Z rather than X that is an unbiased estimator of X . While we can use their results in a black-box manner, they are tailored for the setting where the variance of each Z_{ij} is constant and it is difficult to obtain the exact dependence on this variance in their bounds. In our setting, this variance can be linear in the dimension of the feature vectors, and hence we wish to control the dependence on the variance in the bounds. Thus, we use an algorithm that is similar to the one in Belloni et al. [2016], and provide an analysis for it (in the appendix). As an added bonus, our algorithm results in solving a linear program rather than a conic or general convex program, hence admits a solution that is more computationally efficient.

In the agnostic setting, the computationally efficient algorithm we propose is motivated from (online) supermodular optimization [Natarajan, 1995, Boutsidis et al., 2015, Streeter and Golovin, 2008]. The algorithm is computationally efficient and enjoys sublinear regret under an RIP-like condition, as we will show in Section 3. This result can be contrasted with the known computationally prohibitive algorithms for online

sparse linear regression [Zolghadr et al., 2013, Foster et al., 2016], and the hardness result without RIP [Foster et al., 2015, 2016].

1.3 Notation and Preliminaries

For $d \in \mathbb{N}$, we denote by $[d]$ the set $\{1, 2, \dots, d\}$. For a vector in $x \in \mathbb{R}^d$, denote by $x(i)$ its i -th coordinate. For a subset $S \subseteq [d]$, we use the notation \mathbb{R}^S to indicate the vector space spanned by the coordinate axes indexed by S (i.e. the set of all vectors w supported on the set S). For a vector $x \in \mathbb{R}^d$, denote by $x(S) \in \mathbb{R}^d$ the projection of x on \mathbb{R}^S . That is, the coordinates of $x(S)$ are

$$x(S)(i) = \begin{cases} x(i) & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases} \quad \text{for } i = 1, 2, \dots, d.$$

Let $\langle u, v \rangle = \sum_i u(i) \cdot v(i)$ be the inner product of vectors u and v .

For $p \in [0, \infty]$, the ℓ_p -norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|x\|_p$. For $p \in (0, \infty)$, $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$, $\|x\|_\infty = \max_i |x_i|$, and $\|x\|_0$ is the number of non-zero coordinates of x .

The following definition will play a key role:

Definition 1 (Restricted Isometry Property Candes and Tao [2007]). *Let $\epsilon \in (0, 1)$ and $k \geq 0$. We say that a matrix $X \in \mathbb{R}^{n \times d}$ satisfies restricted isometry property (RIP) with parameters (ϵ, k) if for any $w \in \mathbb{R}^d$ with $\|w\|_0 \leq k$ we have*

$$(1 - \epsilon) \|w\|_2 \leq \frac{1}{\sqrt{n}} \|Xw\|_2 \leq (1 + \epsilon) \|w\|_2.$$

One can show that RIP holds with overwhelming probability if $n = \Omega(\epsilon^{-2} k \log(ed/k))$ and each row of the matrix is sampled independently from an isotropic sub-Gaussian distribution. In the realizable setting, the sub-Gaussian assumption can be relaxed to incorporate heavy tail distribution via the ‘‘small ball’’ analysis introduced in Mendelson [2014], since we only require one-sided lower isometry property.

1.4 Proper Online Sparse Linear Regression

We introduce a variant of online sparse regression (OSLR), which we call *proper online sparse linear regression (POSLR)*. The adjective ‘‘proper’’ is to indicate that the algorithm is required to output a weight vector in each round and its prediction is computed by taking an inner product with the feature vector.

We assume that there is an underlying sequence $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)$ of *labeled examples* in $\mathbb{R}^d \times \mathbb{R}$. In each round $t = 1, 2, \dots, T$, the algorithm behaves according to the following protocol:

1. Choose a vector $w_t \in \mathbb{R}^d$ such that $\|w_t\|_0 \leq k$.
2. Choose $S_t \subseteq [d]$ of size at most k_0 .
3. Observe $x_t(S_t)$ and y_t , and incur loss $(y_t - \langle x_t, w_t \rangle)^2$.

Essentially, the algorithm makes the prediction $\hat{y}_t := \langle x_t, w_t \rangle$ in round t . The regret after T rounds of an algorithm with respect to $w \in \mathbb{R}^d$ is

$$\text{Regret}_T(w) = \sum_{t=1}^T (y_t - \langle x_t, w_t \rangle)^2 - \sum_{t=1}^T (y_t - \langle x_t, w \rangle)^2.$$

The regret after T rounds of an algorithm with respect to the best k -sparse linear regressor is defined as

$$\text{Regret}_T = \max_{w: \|w\|_0 \leq k} \text{Regret}_T(w).$$

Note that any algorithm for POSLR gives rise to an algorithm for OSLR. Namely, if an algorithm for POSLR chooses w_t and S_t , the corresponding algorithm for OSLR queries the coordinates $S_t \cup \{i : w_t(i) \neq 0\}$. The algorithm for OSLR queries at most $k_0 + k$ coordinates and has the same regret as the algorithm for POSLR.

Additionally, POSLR allows parameters settings which do not have corresponding counterparts in OSLR. Namely, we can consider the sparse “full information” setting where $k_0 = d$ and $k \ll d$.

We denote by X_t the $t \times d$ matrix of first t unlabeled samples i.e. rows of X_t are $x_1^T, x_2^T, \dots, x_t^T$. Similarly, we denote by $Y_t \in \mathbb{R}^t$ the vector of first t labels y_1, y_2, \dots, y_t . We use the shorthand notation X, Y for X_T and Y_T respectively.

In order to get computationally efficient algorithms, we assume that for all $t \geq t_0$, the matrix X_t satisfies the restricted isometry condition. The parameter t_0 and RIP parameters k, ϵ will be specified later.

2 Realizable Model

In this section we design an algorithm for POSLR for the realizable model. In this setting we assume that there is a vector $w^* \in \mathbb{R}^d$ such that $\|w^*\|_0 \leq k$ and the sequence of labels y_1, y_2, \dots, y_T is generated according to the linear model

$$y_t = \langle x_t, w^* \rangle + \eta_t, \quad (1)$$

where $\eta_1, \eta_2, \dots, \eta_T$ are independent random variables from $N(0, \sigma^2)$. We assume that the standard deviation σ , or an upper bound of it, is given to the algorithm as input. We assume that $\|w^*\|_1 \leq 1$ and $\|x_t\|_\infty \leq 1$ for all t .

For convenience, we use η to denote the vector $(\eta_1, \eta_2, \dots, \eta_T)$ of noise variables.

2.1 Algorithm

The algorithm maintains an unbiased estimate \widehat{X}_t of the matrix X_t . The rows of \widehat{X}_t are vectors $\widehat{x}_1^T, \widehat{x}_2^T, \dots, \widehat{x}_t^T$ which are unbiased estimates of $x_1^T, x_2^T, \dots, x_t^T$. To construct the estimates, in each round t , the set $S_t \subseteq [d]$ is chosen uniformly at random from the collection of all subsets of $[d]$ of size k_0 . The estimate is

$$\widehat{x}_t = \frac{d}{k_0} \cdot x_t(S_t). \quad (2)$$

To compute the predictions of the algorithm, we consider the linear program

$$\begin{aligned} \text{minimize } \|w\|_1 \text{ s.t. } & \left\| \frac{1}{t} \widehat{X}_t^T (Y_t - \widehat{X}_t w) + \frac{1}{t} \widehat{D}_t w \right\|_\infty \\ & \leq C \sqrt{\frac{d \log(td/\delta)}{tk_0}} \left(\sigma + \frac{d}{k_0} \right). \end{aligned} \quad (3)$$

Here, $C > 0$ is a universal constant, and $\delta \in (0, 1)$ is the allowed failure probability. \widehat{D}_t , defined in equation (5), is a diagonal matrix that offsets the bias on the $\text{diag}(\widehat{X}_t^T \widehat{X}_t)$.

The linear program (3) is called the Dantzig selector. We denote its optimal solution by \widehat{w}_{t+1} . (We define $\widehat{w}_1 = 0$.)

Based on \widehat{w}_t , we construct $\widetilde{w}_t \in \mathbb{R}^d$. Let $|\widehat{w}_t(i_1)| \geq |\widehat{w}_t(i_2)| \geq \dots \geq |\widehat{w}_t(i_d)|$ be the coordinates sorted according to their absolute value, breaking ties according to their index. Let $\widetilde{S}_t = \{i_1, i_2, \dots, i_k\}$ be the top k coordinates. We define \widetilde{w}_t as

$$\widetilde{w}_t = \widehat{w}_t(\widetilde{S}_t). \quad (4)$$

The actual prediction w_t is either zero if $t \leq t_0$ or \widetilde{w}_s for some $s \leq t$ and it gets updated whenever t is a power of 2.

Algorithm 1 Dantzig Selector for POSLR

Require: T, σ, t_0, k, k_0

```
1: for  $t = 1, 2, \dots, T$  do
2:   if  $t \leq t_0$  then
3:     Predict  $w_t = 0$ 
4:   else if  $t$  is a power of 2 then
5:     Let  $\hat{w}_t$  be the solution of linear program (3)
6:     Compute  $\tilde{w}_t$  according to (4)
7:     Predict  $w_t = \tilde{w}_t$ 
8:   else
9:     Predict  $w_t = w_{t-1}$ 
10:  end if
11:  Let  $S_t \subseteq [d]$  be a random subset of size  $k_0$ 
12:  Observe  $x_t(S_t)$  and  $y_t$ 
13:  Construct estimate  $\hat{x}_t$  according to (2)
14:  Append  $\hat{x}_t^T$  to  $\hat{X}_{t-1}$  to form  $\hat{X}_t \in \mathbb{R}^{t \times d}$ 
15: end for
```

The algorithm queries at most $k + k_0$ features each round, and the linear program can be solved in polynomial time using simplex method or interior point method. The algorithm solves the linear program only $\lceil \log_2 T \rceil$ times by using the same vector in the rounds $2^s, \dots, 2^{s+1} - 1$. This lazy update improves both the computational aspects of the algorithm and the regret bound.

2.2 Main Result

The main result in this section provides a logarithmic regret bound under the following assumptions ²

- The feature vectors have the property that for any $t \geq t_0$, the matrix X_t satisfies the RIP condition with $(\frac{1}{5}, 3k)$, with $t_0 = \mathcal{O}(k \log(d) \log(T))$.
- The underlying POSLR online prediction problem has a sparsity budget of k and observation budget k_0 .
- The model is realizable as defined in equation (1) with i.i.d unbiased Gaussian noise with standard deviation $\sigma = \mathcal{O}(1)$.

Theorem 2. *For any $\delta > 0$, with probability at least $1 - \delta$, Algorithm 1 satisfies*

$$\text{Regret}_T = \mathcal{O}(k^2 \log(d/\delta)(d/k_0)^3 \log(T)).$$

The theorem asserts that an $\mathcal{O}(\log T)$ regret bound is efficiently achievable in the realizable setting. Furthermore when $k_0 = \Omega(d)$ the regret scales as $\log(d)$ meaning that we do not necessarily require $T \geq d$ to obtain a meaningful result. We note that the complete expression for arbitrary t_0, σ is given in (13) in the appendix.

The algorithm can be easily understood via the error-in-variable equation

$$\begin{aligned} y_t &= \langle x_t, w^* \rangle + \eta_t, \\ \hat{x}_t &= x_t + \xi_t. \end{aligned}$$

with $\mathbf{E}[\xi_t] = \mathbf{E}[\hat{x}_t - x_t] = 0$, where the expectation is taken over random sampling introduced by the algorithm when performing feature exploration. The learner observes y_t as well as the “noisy” feature vector \hat{x}_t , and aims to recover w^* .

²A more precise statement with the exact dependence on the problem parameters can be found in the appendix.

As mentioned above, we (implicitly) need an unbiased estimator of $X_t^T X_t$. By taking $\widehat{X}_t^T \widehat{X}_t$ it is easy to verify that the off-diagonal entries are indeed unbiased however this is not the case for the diagonal. To this end we define $D_t \in \mathbb{R}^{d \times d}$ as the diagonal matrix compensating for the sampling bias on the diagonal elements of $\widehat{X}_t^T \widehat{X}_t$

$$D_t = \left(\frac{d}{k_0} - 1 \right) \cdot \text{diag} (X_t^T X_t)$$

and the estimated bias from the observed data is

$$\widehat{D}_t = \left(1 - \frac{k_0}{d} \right) \cdot \text{diag} \left(\widehat{X}_t^T \widehat{X}_t \right). \quad (5)$$

Therefore, program (1) can be viewed as Dantzig selector with plug-in unbiased estimates for $X_t^T Y_t$ and $X_t^T X_t$ using limited observed features.

2.3 Sketch of Proof

The main building block in proving Theorem 2 is stated in Lemma 3. It proves that the sequence of solutions \widehat{w}_t converges to the optimal response w^* based on which the signal y_t is created. More accurately, ignoring all second order terms, it shows that $\|\widehat{w}_t - w^*\|_1 \leq \mathcal{O}(1/\sqrt{t})$. In Lemma 4 we show that the same applies for the sparse approximation w_t of \widehat{w}_t . Now, since $\|x_t\|_\infty \leq 1$ we get that the difference between our response $\langle x_t, w_t \rangle$ and the (almost) optimal response $\langle x_t, w^* \rangle$ is bounded by $1/\sqrt{t}$. Given this, a careful calculation of the difference of losses leads to a regret bound w.r.t. w^* . Specifically, an elementary analysis of the loss expression leads to the equality

$$\text{Regret}_T(w^*) = \sum_{t=1}^T 2\eta_t \langle x_t, w^* - w_t \rangle + (\langle x_t, w^* - w_t \rangle)^2$$

A bound on both summands can clearly be expressed in terms of $|\langle x_t, w^* - w_t \rangle| = \mathcal{O}(1/\sqrt{t})$. The right summand requires a martingale concentration bound and the left is trivial. For both we obtain a bound of $\mathcal{O}(\log(T))$.

We are now left with two technicalities. The first is that w^* is not necessarily the empirically optimal response. To this end we provide, in Lemma 16 in the appendix, a constant (independent of T) bound on the regret of w^* compared to the empirical optimum. The second technicality is the fact that we do not solve for \widehat{w}_t in every round, but in exponential gaps. This translates to an added factor of 2 to the bound $\|w_t - w^*\|_1$ that affects only the constants in the $\mathcal{O}(\cdot)$ terms.

Lemma 3 (Estimation Rates). *Assume that the matrix $X_t \in \mathbb{R}^{t \times d}$ satisfies the RIP condition with $(\epsilon, 3k)$ for some $\epsilon < 1/5$. Let $\widehat{w}_{n+1} \in \mathbb{R}^d$ be the optimal solution of program (3). With probability at least $1 - \delta$,*

$$\begin{aligned} \|\widehat{w}_{t+1} - w^*\|_2 &\leq C \cdot \sqrt{\frac{d}{k_0} \cdot \frac{k \log(d/\delta)}{t}} \left(\sigma + \frac{d}{k_0} \right), \\ \|\widehat{w}_{t+1} - w^*\|_1 &\leq C \cdot \sqrt{\frac{d}{k_0} \frac{k^2 \log(d/\delta)}{t}} \left(\sigma + \frac{d}{k_0} \right). \end{aligned}$$

Here $C > 0$ is some universal constant and σ is the standard deviation of the noise.

Note the \widehat{w}_t may not be sparse; it can have many non-zero coordinates that are small in absolute value. However, we take the top k coordinates of \widehat{w}_t in absolute value. Thanks to the Lemma 4 below, we lose only a constant factor $\sqrt{3}$.

Lemma 4. *Let $\widehat{w} \in \mathbb{R}^d$ be an arbitrary vector and let $w^* \in \mathbb{R}^d$ be a k -sparse vector. Let $\widetilde{S} \subseteq [d]$ be the top k coordinates of \widehat{w} in absolute value. Then,*

$$\left\| \widehat{w}(\widetilde{S}) - w^* \right\|_2 \leq \sqrt{3} \|\widehat{w} - w^*\|_2.$$

3 Agnostic Setting

In this section we focus on the agnostic setting, where we don't impose any distributional assumption on the sequence. In this setting, there is no "true" sparse model, but the learner — with limited access to features — is competing with the best k -sparse model defined using full information $\{(x_t, y_t)\}_{t=1}^T$.

As before, we do assume that x_t and y_t are bounded. Without loss of generality, $\|x_t\|_\infty \leq 1$, and $|y_t| \leq 1$ for all t . Once again, without any regularity condition on the design matrix, Foster et al. [2016] have shown that achieving a sub-linear regret $\mathcal{O}(T^{1-\delta})$ is in general computationally hard, for any constant $\delta > 0$ unless $\text{NP} \subseteq \text{BPP}$.

We give an efficient algorithm that achieves sub-linear regret under the assumption that the design matrix of any (sufficiently long) block of consecutive data points has bounded *restricted condition number*, which we define below:

Definition 5 (Restricted Condition Number). *Let $k \in \mathbb{N}$ be a sparsity parameter. The restricted condition number for sparsity k of a matrix $X \in \mathbb{R}^{n \times d}$ is defined as*

$$\sup_{\substack{v, w: \|v\| = \|w\| = 1, \\ \|v\|_0, \|w\|_0 \leq k}} \frac{\|Xv\|}{\|Xw\|}.$$

It is easy to see that if a matrix X satisfies RIP with parameters (ϵ, k) , then its restricted condition number for sparsity k is at most $\frac{1+\epsilon}{1-\epsilon}$. Thus, having bounded restricted condition number is a weaker requirement than RIP.

We now define the *Block Bounded Restricted Condition Number Property* (BBRCNP):

Definition 6 (Block Bounded Restricted Condition Number Property). *Let $\kappa > 0$ and $k \in \mathbb{N}$. A sequence of feature vectors x_1, x_2, \dots, x_T satisfies BBRCNP with parameters (κ, K) if there is a constant t_0 such that for any sequence of consecutive time steps \mathcal{T} with $|\mathcal{T}| \geq t_0$, the restricted condition number for sparsity k of X , the design matrix of the feature vectors x_t for $t \in \mathcal{T}$, is at most κ .*

Note that in the random design setting where x_t , for $t \in [T]$, are isotropic sub-Gaussian vectors, $t_0 = O(\log T + k \log d)$ suffices to satisfy BBRCNP with high probability, where the $O(\cdot)$ notation hides a constant depending on κ .

We assume in this section that the sequence of feature vectors satisfies BBRCNP with parameters (κ, K) for some $K = O(k \log(T))$ to be defined in the course of the analysis.

3.1 Algorithm

The algorithm in the agnostic setting is of distinct nature from that in the stochastic setting. Our algorithm is motivated from literature on maximization of sub-modular set function [Natarajan, 1995, Streeter and Golovin, 2008, Boutsidis et al., 2015]. Though the problem being NP-hard, greedy algorithm on sub-modular maximization provides provable good approximation ratio. Specifically, Streeter and Golovin [2008] considered online optimization of super/sub-modular set functions using expert algorithm as sub-routine. Natarajan [1995], Boutsidis et al. [2015] cast the sparse linear regression as maximization of weakly super-modular function. We will introduce an algorithm that blends various ideas from referred literature, to attack the online sparse regression with limited features.

First, let's introduce the notion of a weakly supermodular function.

Definition 7. *For parameters $k \in \mathbb{N}$ and $\alpha \geq 1$, a set function $g : [d] \rightarrow \mathbb{R}$ is (k, α) -weakly supermodular if for any two sets $S \subseteq T \subseteq [d]$ with $|T| \leq k$, the following two inequalities hold:*

1. **(monotonicity)** $g(T) \leq g(S)$, and
2. **(approximately decreasing marginal gain)**

$$g(S) - g(T) \leq \alpha \sum_{i \in T \setminus S} [g(S) - g(S \cup \{i\})].$$

The definition is slightly stronger than that in Boutsidis et al. [2015]. We will show that sparse linear regression can be viewed as weakly supermodular minimization in Definition 7 once the design matrix has bounded restricted condition number.

Now we outline the algorithm (see Algorithm 2). We divide the rounds $1, 2, \dots, T$ into mini-batches of size B each (so there are T/B such batches). The b -th batch thus consists of the examples (x_t, y_t) for $t \in \mathcal{T}_b := \{(b-1)B+1, (b-1)B+2, \dots, bB\}$. Within the b -th batch, our algorithm queries the same subset of features of size at most k_0 .

The algorithm consists of few key steps. First, one can show that under BBRCNP, as long as B is large enough, the loss within batch b defines a weakly supermodular set function

$$g_t(S) = \frac{1}{B} \inf_{w \in \mathbb{R}^S} \sum_{t \in \mathcal{T}_b} (y_t - \langle x_t, w \rangle)^2.$$

Therefore, we can formulate the original online sparse regression problem into online weakly supermodular minimization problem. For the latter problem, we develop an online greedy algorithm along the lines of [Streeter and Golovin, 2008]. We employ $k_1 = \mathcal{O}^*(k)$ budgeted experts algorithms [Amin et al., 2015], denoted BEXP, with budget parameter³ $\frac{k_0}{k_1}$. The precise characteristics of BEXP are given in Theorem 8 (adapted from Theorem 2 in [Amin et al., 2015]).

Theorem 8. *For the problem of prediction from expert advice, let there be d experts, and let $k \in [d]$ be a budget parameter. In each prediction round t , the BEXP algorithm chooses an expert j_t and a set of experts U_t containing j_t of size at most k , obtains as feedback the losses of all the experts in U_t , suffers the loss of expert j_t , and guarantees an expected regret bound of $2\sqrt{\frac{d \log(d)}{k}} T$ over T prediction rounds.*

At the beginning of each mini-batch b , the BEXP algorithms are run. Each BEXP algorithm outputs a set of coordinates of size $\frac{k_0}{k_1}$ as well as a special coordinate in that set. The union of all of these sets is then used as the set of features to query throughout the subsequent mini-batch. Within the mini-batch, the algorithm runs the standard Vovk-Azoury-Warmuth algorithm for linear prediction with square loss restricted to set of special coordinates output by all the BEXP algorithms.

At the end of the mini-batch, every BEXP algorithm is provided carefully constructed losses for each coordinate that was output as feedback. These losses ensure that the set of special coordinates chosen by the BEXP algorithms mimic the greedy algorithm for weakly supermodular minimization.

3.2 Main Result

In this section, we will show that Algorithm 2 achieves sublinear regret under BBRCNP.

Theorem 9. *Suppose the sequence of feature vectors satisfies BBRCNP with parameters $(\kappa, k_1 + k)$ for $k_1 = \frac{1}{3}\kappa^2 k \log(T)$, and assume that T is large enough so that $t_0 \leq (\frac{k_0 T}{\kappa^2 dk})^{1/3}$. Then if Algorithm 2 is run with parameters $B = (\frac{k_0 T}{\kappa^2 dk})^{1/3}$ and k_1 as specified above, its expected regret is at most $\tilde{O}((\frac{\kappa^8 dk^4}{k_0})^{1/3} T^{2/3})$.*

Proof. The proof relies on a number of lemmas whose proofs can be found in the appendix. We begin with the connection between sparse linear regression, weakly supermodular function and RIP, formally stated in Lemma 10. This lemma is a direct consequence of Lemma 5 in [Boutsidis et al., 2015].

Lemma 10. *Consider a sequence of examples $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ for $t = 1, 2, \dots, B$, and let X be the design matrix for the sequence. Consider the set function associated with least squares optimization:*

$$g(S) = \inf_{w \in \mathbb{R}^S} \frac{1}{B} \sum_{t=1}^B (y_t - \langle x_t, w \rangle)^2.$$

Suppose the restricted condition number of X for sparsity k is bounded by κ . Then $g(S)$ is (k, κ^2) -weakly supermodular.

³We assume, for convenience, that k_0 is divisible by k_1 .

Algorithm 2 Online Greedy Algorithm for POSLR

Require: Mini-batch size B , sparsity parameters k_0 and k_1

- 1: Set up k_1 budgeted prediction algorithms $\text{BEXP}^{(i)}$ for $i \in [k_1]$, each using the coordinates in $[d]$ as “experts” with a per-round budget of $\frac{k_0}{k_1}$.
- 2: **for** $b = 1, 2, \dots, T/B$ **do**
- 3: For each $i \in [k_1]$, obtain a coordinate $j_b^{(i)}$ and subset of coordinates $U_b^{(i)}$ from $\text{BEXP}^{(i)}$ such that $j_b^{(i)} \in U_b^{(i)}$.
- 4: Define $V_b^{(0)} = \emptyset$ and for each $i \in [k_1]$ define $V_b^{(i)} = \{j_b^{(i')} \mid i' \leq i\}$.
- 5: Set up the Vovk-Azoury-Warmuth (VAW) algorithm for predicting using the features in $V_b^{(k_1)}$.
- 6: **for** $t \in \mathcal{T}_b$ **do**
- 7: Set $S_t = \bigcup_{i \in [k_1]} U_b^{(i)}$, obtain $x_t(S_t)$, and pass $x_t(V_b^{(k_1)})$ to VAW.
- 8: Set w_t to be the weight vector output by VAW.
- 9: Obtain the true label y_t and pass it to VAW.
- 10: **end for**
- 11: Define the function

$$g_b(S) = \frac{1}{B} \inf_{w \in \mathbb{R}^S} \sum_{t \in \mathcal{T}_b} (y_t - \langle x_t, w \rangle)^2. \quad (6)$$

- 12: For each $j \in U_b^{(i)}$, compute $g_b(V_b^{(i-1)} \cup \{j\})$ and pass it to $\text{BEXP}^{(i)}$ as the loss for expert j .
 - 13: **end for**
-

Even though minimization of weakly supermodular functions is NP-hard, the greedy algorithm provides a good approximation, as shown in the next lemma.

Lemma 11. *Consider a (k, α) -weakly supermodular set function $g(\cdot)$. Let $j^* := \arg \min_j g(\{j\})$. Then, for any subset V of size at most k , we have*

$$g(\{j^*\}) - g(V) \leq \left(1 - \frac{1}{\alpha|V|}\right) [g(\emptyset) - g(V)].$$

The BEXP algorithms essentially implement the greedy algorithm in an online fashion. Using the properties of the BEXP algorithm, we have the following regret guarantee:

Lemma 12. *Suppose the sequence of feature vectors satisfies BBRCNP with parameters $(\epsilon, k_1 + k)$. Then for any set V of coordinates of size at most k , we have*

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

Finally, within every mini-batch, the VAW algorithm guarantees the following regret bound, an immediate consequence of Theorem 11.8 in Cesa-Bianchi and Lugosi [2006]:

Lemma 13. *Within every batch b , the VAW algorithm generates weight vectors w_t for $t \in \mathcal{T}_b$ such that*

$$\sum_{t \in \mathcal{T}_b} (y_t - \langle x_t, w_t \rangle)^2 - Bg_b(V_b^{(k_1)}) \leq O(k_1 \log(B)).$$

We can now prove Theorem 9. Combining the bounds of lemma 12 and 13, we conclude that for any subset of coordinates V of size at most k , we have

$$\mathbf{E} \left[\sum_{t=1}^T (y_t - \langle x_t, w_t \rangle)^2 \right] \tag{7}$$

$$\leq \sum_{b=1}^{T/B} B g_b(V) + B \left(1 - \frac{1}{\kappa^2 |V|}\right)^{k_1} [g_b(\emptyset) - g_b(V)] \tag{8}$$

$$+ O \left(\kappa^2 k \sqrt{\frac{d k_1 \log(d) B T}{k_0}} + \frac{T}{B} k_1 \log(B) \right). \tag{9}$$

Finally, note that

$$\sum_{b=1}^{T/B} B g_b(V) \leq \inf_{w \in \mathbb{R}^V} \sum_{t=1}^T (y_t - \langle x_t, w \rangle)^2,$$

and

$$\sum_{b=1}^{T/B} B \left(1 - \frac{1}{\kappa^2 |V|}\right)^{k_1} [g_b(\emptyset) - g_b(V)] \leq T \cdot \exp\left(-\frac{k_1}{\kappa^2 k}\right),$$

because $g_b(\emptyset) \leq 1$. Using these bounds in (9), and plugging in the specified values of B and k_1 , we get the stated regret bound. \square

4 Conclusions and Future Work

In this paper, we gave computationally efficient algorithms for the online sparse linear regression problem under the assumption that the design matrices of the feature vectors satisfy RIP-type properties. Since the problem is hard without any assumptions, our work is the first one to show that assumptions that are similar to the ones used to sparse recovery in the batch setting yield tractability in the online setting as well.

Several open questions remain in this line of work and will be the basis for future work. Is it possible to improve the regret bound in the agnostic setting? Can we give matching lower bounds on the regret in various settings? Is it possible to relax the RIP assumption on the design matrices and still have efficient algorithms? Some obvious weakenings of the RIP assumption we have made don't yield tractability. For example, simply assuming that the final matrix X_T satisfies RIP rather than every intermediate matrix X_t for large enough t is not sufficient; a simple tweak to the lower bound construction of Foster et al. [2016] shows this. This tweak consists of simply padding the construction with enough dummy examples which are well-conditioned enough to overcome the ill-conditioning of the original construction so that RIP is satisfied by X_T . We note however that in the realizable setting, our analysis can be easily adapted to work under weaker conditions such as irrepresentability [Zhao and Yu, 2006, Javanmard and Montanari, 2013].

References

- Kareem Amin, Satyen Kale, Gerald Tesauro, and Deepak S. Turaga. Budgeted prediction with expert advice. In *AAAI*, pages 2490–2496, 2015.
- Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B. Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. ISSN 1467-9868.
- Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pages 1705–1732, 2009.

- Christos Boutsidis, Edo Liberty, and Maxim Sviridenko. Greedy minimization of weakly supermodular set functions. *arXiv preprint arXiv:1502.06528*, 2015.
- Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n . *The Annals of Statistics*, pages 2313–2351, 2007.
- Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.
- Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.
- Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 12(Oct):2857–2878, 2011.
- Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *COLT*, pages 696–709, 2015.
- Dean Foster, Satyen Kale, and Howard Karloff. Online sparse linear regression. In *COLT*, 2016.
- Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.
- Elad Hazan and Tomer Koren. Linear regression with limited observation. In *ICML*, 2012.
- Adel Javanmard and Andrea Montanari. Model selection for high-dimensional regression under the generalized irrepresentability condition. In *NIPS*, pages 3012–3020, 2013.
- Satyen Kale. Open problem: Efficient online sparse regression. In *COLT*, pages 1299–1301, 2014.
- Doron Kukliansky and Ohad Shamir. Attribute efficient linear regression with distribution-dependent sampling. In *ICML*, pages 153–161, 2015.
- Shahar Mendelson. Learning without concentration. In *COLT*, pages 25–39, 2014.
- Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.
- Mathieu Rosenbaum and Alexandre B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.
- Matthew J. Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In *NIPS*, pages 1577–1584, 2008.
- Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.
- Navid Zolghadr, Gábor Bartók, Russell Greiner, András György, and Csaba Szepesvári. Online learning with costly features and labels. In *NIPS*, pages 1241–1249, 2013.

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, \dots, 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 $\{i_{(j-1)k+1}, \dots, i_{jk}\}$. 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 $c \in \mathbb{R}^{S_{01}}$,

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

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

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

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

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

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

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

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 and Tao [2005] stating that for any S', S'' with $|S'| + |S''| \leq 3k$ it holds that

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

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

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

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

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

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 $\Delta(i) \leq (1/k) \|\Delta(S_{j-1})\|_1$. Hence

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

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

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

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 , $\hat{w}_{S^c} = \Delta(S^c)$ hence

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

Now, the optimality of \hat{w} implies $\|\hat{w}\|_1 \leq \|w^*\|_1$, 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

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

Rearranging we conclude that

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

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

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

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 $\|\Delta(S_{01}^c)\|_2^2 = \sum_{j \geq k+1} \alpha_j^2$ 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. \square

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,

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

Taking square root finishes the proof. \square

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*

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

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 $X = X_t$, $\hat{X} = X_t$ and $\hat{D} = \hat{D}_t$, and also let $\eta = (\eta_1, \eta_2, \dots, \eta_t)$ 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

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

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

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

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

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

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

$$\frac{1}{t} \left\| \hat{X}^T \eta \right\|_\infty \leq C \sigma \sqrt{\frac{d \log(d/\delta)}{t k_0}}.$$

Combining the inequalities so far provides

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

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 $\Delta := \hat{w} - w^*$, we have

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

According to Lemma 15 we have

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

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

$$\left\| \left(\frac{1}{t}(\hat{X} - X)^T (\hat{X} - X) - \frac{1}{t} \hat{D} \right) \Delta \right\|_\infty \leq 2C \sqrt{\frac{d^3 \log(d/\delta)}{tk_0^3}} \cdot \|w^*\|_1$$

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

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

and reach the final bound. \square

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

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

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) $(\hat{x}_s(i) - x_s(i))^2 - \frac{1}{t} D_{ii}$ 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

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

For $i = j$, we have

$$\begin{aligned} \mathbf{E} \left[\left((\hat{x}_s(i) - x_s(i))^2 - \frac{1}{t} D_{ii} \right)^2 \right] &\leq \mathbf{E} [(\hat{x}_s(i) - x_s(i))^4] \leq (d/k_0)^3 \\ (\hat{x}_s(i) - x_s(i))^2 - \frac{1}{t} D_{ii} &\leq (d/k_0)^2, \quad \mathbf{E} \left[(\hat{x}_s(i) - x_s(i))^2 - \frac{1}{t} D_{ii} \right] = 0 \end{aligned}$$

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

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

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

$$\left| \frac{1}{t} \sum_{s=1}^t (\hat{x}_s(i) - x_s(i))^2 - \frac{1}{t} D_{ii} \right| \leq \mathcal{O} \left(\frac{\log(d/\delta) d^2}{tk_0^2} + \sqrt{\frac{\log(d/\delta) d^3}{tk_0^3}} \right).$$

Similarly we have $\frac{1}{t}(\hat{D}_{ii} - D_{ii}) \leq \mathcal{O} \left(\frac{\log(d/\delta) d^2}{tk_0^2} + \sqrt{\frac{\log(d/\delta) d^3}{tk_0^3}} \right)$.

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 $1/\sqrt{t}$. \square

Proof of Theorem 2. By Lemma 3,

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

We have

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

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 $1 - \delta$,

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

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

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

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 $\eta_1 \langle x_1, w^* \rangle, \eta_2 \langle x_2, w^* \rangle, \dots, \eta_{t_0} \langle x_{t_0}, w^* \rangle$ 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 $\text{Regret}_T - \text{Regret}_T(w^*)$, gives

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

□

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

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 $\|X_SX_S^\dagger\eta\|^2$ 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 $\|X_SX_S^\dagger\eta\|^2/\sigma^2$ is distributed according to the χ_{2k}^2 distribution. The tail bounds of this distribution suggest that

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

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

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

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 $g(S) - g(T)$.

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

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 $\|(Z_{T \setminus S}^T)^\dagger\| = \|Z_{T \setminus S}^\dagger\| \leq \|X_T^\dagger\|$. 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 $\zeta_i \leq \|x_i\| \leq 1$ we must have $\|w'\| \geq \|w\|$. 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

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

It follows that

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

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:

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

□

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 $X_j = X e_j$ 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 $\alpha = \max_{S: |S| \leq k} \|\bar{X}_S^\dagger\|_2^2$. 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 $\frac{1}{\|\bar{X}_S w\|} = \frac{\|X e_j\|}{\|X w'\|}$, 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 $\frac{\|X e_j\|}{\|X w'\|} \leq \kappa$. Since this bound holds for any subset S of size at most k , we conclude that $\alpha \leq \kappa^2$. \square

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

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

Rearranging, we get the claimed bounds. \square

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 $|T| \leq k - |U|$, we have

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

\square

Proof of Lemma 12. For $i \in \{0, 1, \dots, k_1\}$, 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 $\text{BEXP}^{(i)}$. 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 $0 \leq g_b(S) \leq \frac{1}{B} \sum_{t \in \mathcal{T}_b} y_t^2 \leq 1$. Thus, the regret guarantee of BEXP (Theorem 8) implies that for any $i \in [k_1]$ and any $j \in [d]$, we have

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

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

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

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 $|V_b^{(i)}| \leq k_1$).

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 ,

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

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

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

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

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

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

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

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