A Polynomial-time Algorithm for Online Sparse Linear Regression with Improved Regret Bound under Weaker Conditions

Junfan Li LIJUNFAN@HIT.EDU.CN
Harbin Institute of Technology (Shenzhen)
Shizhong Liao SZLIAO@TJU.EDU.CN
Tianjin University
Zenglin Xu * ZENGLIN@GMAIL.COM
Fudan University and Shanghai Academy of AI for Science
Liqiang Nie * NIELIQIANG@GMAIL.COM
Harbin Institute of Technology (Shenzhen)

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In this paper, we study the problem of online sparse linear regression (OSLR) where the algorithms are restricted to accessing only k out of d attributes per instance for prediction, which was proved to be NP-hard. Previous work gave polynomial-time algorithms assuming the data matrix satisfies the linear independence of features, the compatibility condition, or the restricted isometry property. We introduce a new polynomial-time algorithm, which significantly improves previous regret bounds (Ito et al., 2017) under the compatibility condition that is weaker than the other two assumptions. The improvements benefit from a tighter convergence rate of the ℓ_1-norm error of our estimators. Our algorithm leverages the well-studied Dantzig Selector, but importantly with several novel techniques, including an algorithm-dependent sampling scheme for estimating the covariance matrix, an adaptive parameter tuning scheme, and a batching online Newton step with careful initializations. We also give novel and non-trivial analyses, including an induction method for analyzing the ℓ_1-norm error, careful analyses on the covariance of non-independent random variables, and a decomposition on the regret. We further extend our algorithm to OSLR with additional observations where the algorithms can observe additional k_0 attributes after each prediction, and improve previous regret bounds (Kale et al., 2017; Ito et al., 2017).

Keywords: Sparse linear regression, online learning, compatibility condition, Dantzig Selector

1. Introduction

For most online prediction tasks, algorithms are assumed to observe all of the attributes of an instance $\mathbf{x}_t \in \mathbb{R}^d$ at each round $t = 1, 2, \ldots, T$. However, the assumption is hard to be satisfied in some real-world scenarios due to various constraints, such as computational constraint, human labors and privacy constraint (Hazan and Koren, 2012; Jain et al., 2012; Zolghadr et al., 2013; Murata and Suzuki, 2018). In the task of medical diagnosis of a disease (Cesa-Bianchi et al., 2010), $\mathbf{x}_t$ contains the results of a large number of medical tests. However, many patients can only pay the cost for several medical tests. Thus $\mathbf{x}_t$ must be sparse. Another example is personalized recommendation (Jain et al., 2012). Due to the privacy constraint, search engines are not allowed to use sensitive attributes of users, such as gender, age, job and so on. $\mathbf{x}_t$ is also sparse. The algorithms typically make predictions using only limited attributes.

* Corresponding author.

Kale (2014) first formulated online prediction problems with limited access to attributes as online sparse linear regression (OSLR). At each round t, an adversary gives an instance $\mathbf{x}_t$ to a learner. The learner can only observe k, $k < d$, attributes of $\mathbf{x}_t$ at most, and outputs a prediction $\hat{y}_t$. Then the adversary gives the true output y_t. The learner suffers a loss $(\hat{y}_t - y_t)^2$. The learner aims to minimize her cumulative losses over T rounds. Typically, we compare the cumulative losses of the learner with that of any k-sparse competitor and define the regret as follows,

$$\forall \mathbf{w} \in \left\{ \mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_0 \leq k \right\}, \quad \mathrm{Reg}(\mathbf{w}) = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2 - \sum_{t=1}^{T} \left(\mathbf{w}^\top \mathbf{x}_t - y_t \right)^2. \tag{1}$$

The primal goal is to develop an algorithm that runs in time $O(\mathrm{poly}(T, k, d))$ per-iteration, and enjoys a (an expected) regret of order $O(\mathrm{poly}(d, k) \cdot o(T))$. There is also a relaxation of OSLR denoted by (k, k_0, d)-OSLR, in which the algorithms are allowed to observe additional $k_0 = O(k \ln d)$ attributes after each prediction.

The offline problems related to OSLR and (k, k_0, d)-OSLR concern the approximation of the k-sparse solution in linear systems that has been proven to be NP-hard by a reduction from set cover problem (Natarajan, 1995; Foster et al., 2015). By a similar approach, it was shown that both OSLR and (k, k_0, d)-OSLR are also NP-hard (Foster et al., 2016). Specifically, there is no algorithm running in time $O(\mathrm{poly}(T, k, d))$ per-iteration and having an expected regret of $O(\mathrm{poly}(d) \cdot T^{1-\delta})$ for any constant $\delta > 0$ unless $\mathbf{NP} \subseteq \mathbf{BPP}$. They also proposed an inefficient algorithm for (k, k_0, d)-OSLR that enjoys an expected regret of $O(\frac{d^2}{k_0^2}\sqrt{kT \ln d})$ at a $O((k + k_0)\binom{d}{k})$ space and per-round time complexity. Given the computational hardness result, it is necessary to further restrict the problem with additional regularity assumptions. Certain assumptions on the data matrix are adequate for approximating the sparse solutions of linear systems, such as the restricted isometry property (RIP) (Candès and Tao, 2005), the compatibility condition (van de Geer, 2007), the restricted eigenvalues condition (Bickel et al., 2009), the restricted strong convexity (RSC) and smoothness of the loss function (Murata and Suzuki, 2018), and so on. Kale et al. (2017) proposed the first algorithm for (k, k_0, d)-OSLR (under a different name POSLR) based on the Dantzig Selector (Candès and Tao, 2007) that runs in time $O(\mathrm{poly}(d))$ per-iteration and enjoys a high-probability regret of $O(\frac{k^2 d^3 \ln T}{k_0^3} \ln \frac{Td}{\delta})$ under RIP. Ito et al. (2017) independently proposed algorithms for OSLR and (k, k_0, d)-OSLR, all of which run in time $O(d)$ per-iteration and enjoy an expected regret of $O(\mathrm{poly}(d, k) + \mathrm{poly}(d, k)\sqrt{T})$ under the linear independence of features or compatibility condition.

However, previous computationally efficient algorithms either exhibit undesirable reliance on problem-dependent parameters like T, d, $\min_{i \in S} |w_i^*| < 1$, and numerical factor, or require stringent assumptions, in which $\mathbf{w}^* = (w_1^*, \ldots, w_d^*)$ is a k-sparse vector with support set S such that $y_t = \langle \mathbf{x}_t, \mathbf{w}^* \rangle + \eta_t$ where η_t is a noise. Specifically, for (k, k_0, d)-OSLR, the regret bound of the algorithm (Kale et al., 2017) depends on $O(d^3)$ under the stringent RIP, and the regret bound of the algorithm (Ito et al., 2017) depends on $O(\sqrt{T})$ under the stronger linear independent features condition. For OSLR, the regret bounds of the two algorithms (Ito et al., 2017) depend on d^8 [1], $\min_{i \in S} |w_i^*|^{-7}$ or a large constant factor $128^2 \cdot 36^8$ under the linear independent features or the compatibility condition. Table 1 and Table 2 show previous regret bounds. It is natural to ask

1. There are typos in original paper (Ito et al., 2017). The correct regret bounds are $O(\frac{d^8}{k^8})$, not $O(\frac{d^{16}}{k^{16}})$ as erroneously stated on Page 2 of the original paper.

Algorithm	Regret bound w.r.t. $\mathbf{w}^*$	Per-round time	Assumptions
Ito et al. (2017)	$8\sqrt{kT} + \frac{8192^2 \cdot d^8}{\sigma_d^8 k^7 h(\mathbf{w}^*)^7} + O(1)$	$O(d)$	(a), (b), (c), (g), (i)
Ito et al. (2017)	$8\sqrt{kT} + \frac{128^2 \cdot 36^8 \cdot d^8}{\delta_S^8 k^3 h(\mathbf{w}^*)^7} + O(1)$	$O(d)$	(a), (b), (d), (g), (i)
DS-OSLRC (Ours)	$4\sqrt{T} + \frac{102^4 k^2 d^2}{\delta_S^8 h(\mathbf{w}^*)^2} \ln^2 \frac{dT}{\delta} + O(1)$	$O\left(\frac{\mathrm{LP}_d}{\sqrt{T}} + k^2\right)$	(a), (b), (d), (f)

Table 1: Comparison of the algorithms for OSLR. $\delta \in (0, 1)$ is a probability parameter, $h(\mathbf{w}^*) = \min_{i \in S} |w_i^*| < 1$, $\sigma_d \leq \delta_S$. $\mathrm{LP}_d = O(\mathrm{poly}(d))$ is the time complexity of solving a linear programming with d constraints and d variables. Assumptions: (a) realizable, (b) sparsity, (c) linear independence of features, (d) compatibility condition, (g) bounded noise, (f) Gaussian noise, and (i) i.i.d. instances. **Assumption (c) is stronger than (d).**

Question *Whether there are polynomial-time algorithms for OSLR with regret bounds that exhibit improved dependence on T, d, $\min_{i \in S} |w_i^*|$ and numerical factor, and for (k, k_0, d)-OSLR with regret bounds that exhibit improved dependence on T and d under mild assumptions?*

1.1. Main Results

In this paper, we will answer the question affirmatively. We first propose a new polynomial-time algorithm for OSLR, named DS-OSLRC, and then extend the algorithm to (k, k_0, d)-OSLR, named DS-POSLRC. Our algorithms enjoy much better regret bounds under the realizable assumption, that is, there is a true k-sparse vector, and the compatibility condition that is less restrictive than RIP and the linear independent features condition. Our main results are summarized in Table 1 and Table 2.

- DS-OSLRC outputs an estimator $\hat{\mathbf{w}}_s$ satisfying $\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1 = \tilde{O}(\sqrt{kd/s})$ for sufficiently large values of s, improving previous convergence rates (Ito et al., 2017; Kale et al., 2017).
- For OSLR, DS-OSLRC enjoys a $4\sqrt{T} + \frac{102^4 k^2 d^2}{\delta_S^8 h(\mathbf{w}^*)^2} \ln^2 \frac{dT}{\delta} + O(1)$ high-probability regret bound, in which $h(\mathbf{w}^*) = \min_{i \in S} |w_i^*| < 1$ and $O(1)$ hides some lower order terms. The averaging per-round time complexity is $O(\frac{\mathrm{LP}_d}{\sqrt{T}} + k^2)$, in which $\mathrm{LP}_d = O(\mathrm{poly}(d))$. We significantly improve the regret bounds by Ito et al. (2017) in terms of d, T, $h(\mathbf{w}^*)$ and numerical factor, under weaker or the same assumptions. Besides, the regret bounds in Ito et al. (2017) hold in expectation, which are weaker than our high-probability regret bound.
- For (k, k_0, d)-OSLR, DS-POSLRC enjoys a $O(\frac{k^2 d^2}{\delta_S^4 k_0^2} \ln \frac{dT}{\delta} + \frac{k^2 d}{\delta_S^4 k_0} \ln(T) \ln \frac{dT}{\delta})$ high-probability regret bound. The per-round time complexity is $O(\mathrm{LP}_d)$. We improve the regret bound by Kale et al. (2017) by a factor of $O(\min\{\frac{d^2}{k_0^2}, \frac{d}{k_0} \ln T\})$, and significantly improve the regret bound by Ito et al. (2017) in terms of T, under weaker assumptions.

1.2. Technical Challenges and Contributions

DS-OSLRC builds upon the well-studied Dantzig Selector (Candès and Tao, 2007). Despite this foundation, achieving such significant improvements on the regret bounds necessitates more novel methodologies in algorithm design and regret analyses. We explain as follows.

Algorithm	Regret bound w.r.t. $\mathbf{w}^*$	Per-round time	Assumptions	k_0
Foster et al. (2016)	$O\left(\frac{d^2}{k_0^2}\sqrt{kT\ln d}\right)$	$O(d^k)$	-	≥ 2
Ito et al. (2017)	$O\left(\frac{d}{\sigma_d^2 k_0}\sqrt{T}\right)$	$O(d)$	(a), (b), (c), (g), (i)	≥ 2
Kale et al. (2017)	$O\left(\frac{k^2d^3}{k_0^3}\ln(T)\ln\frac{dT}{\delta}\right)$	$O(\frac{\mathrm{LP}_d}{T/\log T}+d)$	(a), (b), (e), (f)	≥ 2
DS-POSLRC	$O\left(\frac{k^2 d}{\delta_S^4 k_0}(\frac{d}{k_0}+\ln T)\ln\frac{dT}{\delta}\right)$	$O\left(\mathrm{LP}_d\right)$	(a), (b), (d), (f)	≥ 3

Table 2: Comparison of the algorithms for (k, k_0, d)-OSLR. Assumption (e) is RIP, and the others follow Table 1. **Assumption (c) is stronger than (e). (e) is stronger than (d)**.

The improvements on problem-dependent parameters, such as $d, T, h(\mathbf{w}^*)$, stem from a new algorithm-dependent sampling scheme, and a batching online Newton step (ONS) (Hazan et al., 2007) with careful initializations. At some exploration round $s > 1$, let $\hat{\mathbf{w}}_s$ be the estimator of $\mathbf{w}^*$. Previous algorithms (Kale et al., 2017; Ito et al., 2017) uniformly sample k or k_0 attributes from $\mathbf{x}_s$ to estimate $\mathbf{x}_s$ and $\mathbf{x}_s\mathbf{x}_s^\top$. Whereas our sampling scheme uses $\hat{\mathbf{w}}_{s-1}$ to construct sampling probability and simultaneously combines uniformly sampling without replacement. What's more, it requires novel analyses on $\|\Delta_s(S)\|_1$, in which $\Delta_s(S) = \hat{\mathbf{w}}_s(S) - \mathbf{w}^*$ and $\hat{\mathbf{w}}_s(S)$ extracts the elements of $\hat{\mathbf{w}}_s$ restricted to $S \subseteq [d]$. We will prove that $\|\Delta_s(S)\|_1$ satisfies the following inequality

$$\frac{\delta_S^2\|\Delta_s(S)\|_1}{k} \leq c_1\frac{(d-1)(d-2)}{(k-1)(k-2)s}\ln\frac{d}{\delta} + c_2\sqrt{\frac{d-1}{s(k-1)}\ln\frac{d}{\delta}} + \mu_s + \frac{c_3}{s}\sqrt{\frac{(d-1)(d-2)}{(k-1)(k-2)}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}} + c_4\sqrt{\frac{(d-1)(d-2)}{(k-1)(k-2)s}\ln\frac{d^2}{\delta}}\|\Delta_s(S)\|_1, \tag{2}$$

where $\delta_S, c_1, c_2, c_3, c_4$ are constants and μ_s is an estimator of $\sqrt{\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}}$. We will use a novel induction method to solve the inequality and prove that $\|\Delta_s(S)\|_1 = \tilde{O}(\sqrt{kd/s})$ for sufficiently large s, making it possible to refine the dependence on d and $h(\mathbf{w}^*)$. We further refine the dependence on T by using ONS to update parameters throughout each epoch $\mathcal{T}_s \subseteq [T]$. Through novel analyses, DS-OSLRC enjoys a regret of $O(\sqrt{T} + \frac{k^2d^2}{h(\mathbf{w}^*)^2}\ln^2(dT))$.

The algorithm-dependent sampling scheme also poses a technical challenge on the algorithm. To be specific, at each exploration round s, there is a threshold $\gamma_s > 0$ related the Dantzig Selector, in which γ_s depends on $\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2$. As $\mathbf{w}^*$ is unknown, it is impossible to use the optimal γ_s. A similar issue exists in the online linearized LASSO algorithm (Yang et al., 2023), in which the regularization parameter depends on $\|\Delta_{s-1}\|_1$. In fact, it is enough to construct $\hat{\gamma}_s \geq c\gamma_s$, where c is a constant. To this end, we use a "guess-then-verification" technique. Specifically, if γ_s is known, then we can set an exact value for μ_s in (2) and solve the inequality. In this way, we obtain the ideal convergence rate that allows us to infer $\hat{\gamma}_s$. Then we use an induction method to verify the correctness, and solve (2) for obtaining the real convergence rate.

The improvement on the constant factor also stems from a tighter convergence rate of $\|\Delta_s(S)\|_1$, and requires novel analyses, including careful analyses on the covariance of non-independent random variables, and a decomposition on the regret. Since DS-OSLRC samples k attributes without

replacement, the observed attributes are not independent. It is necessary to control the covariances. Regarding the regret analysis, we decompose the regret in each epoch $\mathcal{T}_s$ into two components,

$$\underbrace{\sum_{t\in\mathcal{T}_s}[\ell(\langle\hat{\mathbf{w}}_s(S_s),\mathbf{x}_t\rangle,y_t)-\ell(\langle\mathbf{w}_s,\mathbf{x}_t\rangle,y_t)]}_{\mathrm{R}_1}+\underbrace{\sum_{t\in\mathcal{T}_s}[\ell(\langle\mathbf{w}_s,\mathbf{x}_t\rangle,y_t)-\ell(\langle\mathbf{w}^*,\mathbf{x}_t\rangle,y_t)]}_{\mathrm{R}_2},$$

in which $\mathrm{Supp}(\mathbf{w}_s)\subset S_s$. R_1 can be bounded by the regret of ONS. Moreover, the constant factor associated with R_1 is quite small. R_2 depends on $\|\mathbf{w}_s-\mathbf{w}^*\|_1$. We note that $\mathbf{w}_s$ can be selected arbitrarily. We will use $\mathbf{w}_s=\mathbf{w}^*(S_s\cap S)$, rather than the obvious but sub-optimal $\hat{\mathbf{w}}_s(S_s)$. It can be proved $\|\mathbf{w}^*(S_s\cap S)-\mathbf{w}^*\|_1\leq\|\hat{\mathbf{w}}_s(S_s)-\mathbf{w}^*\|_1$, thereby further reducing the numerical factor.

1.3. Related Work

Yang et al. (2023) formulated another variant of online sparse linear regression where the algorithms observe all attributes of each $\mathbf{x}_s$, $s=1,\ldots,T$, and aim to compute a sequence of estimators $\{\hat{\mathbf{w}}_s\}_{s=1}^T$ approximating $\mathbf{w}^*$, and proposed an online linearized LASSO algorithm, named OLin-LASSO. The algorithm was proved to achieve the optimal $\tilde{O}(\sqrt{k/s})$ ℓ_2-norm error bound under the RSC condition (Murata and Suzuki, 2018). RSC is stronger than the restricted eigenvalues and the compatibility condition, but is weaker than RIP. The DA-GL algorithm (Yang et al., 2010), which uses dual averaging to solve group LASSO, can also be applied to solve this variant. Without additional assumptions, DA-GL runs in time $O(d)$ per-iteration and returns a solution whose ℓ_2-norm error is bounded by a constant. The I-LAMM algorithm (Fan et al., 2018) first computes a good initialization $\hat{\mathbf{w}}_0$ on a batch of initial data, and then iteratively computes a sequence of better estimators. I-LAMM is also optimal under the restricted eigenvalues condition. The three algorithms are clearly unsuitable for OSLR and (k,k_0,d)-OSLR, as they require the complete information of each instance.

Another related problem is linear regression with limited observations (LRLO) (Cesa-Bianchi et al., 2010, 2011; Hazan and Koren, 2012) where the algorithms only observe k attributes of each instance at training phase, but aim to learn a linear model with low excess risk. Murata and Suzuki (2018) studied a harder variant of LRLO where the algorithms only observe k attributes of each instance at both training and test phase, and proposed an exploration-then-exploitation algorithm with improved sample complexity under RSC. Since the algorithms are not required to give predictions for training instances, the two problems are inherently less complex than OSLR. In other words, any algorithm for OSLR is applicable for the two problems. However, the reverse is not true.

2. Notations and Preliminaries

2.1. Notations

For any vector $\mathbf{x}\in\mathbb{R}^d$, denote by x_i its i-th coordinate. For any matrix $\mathbf{X}\in\mathbb{R}^{m\times n}$, denote by $X_{i,j}$ the element on the i-th row and j-th column. For any $d\in\mathbb{N}$, let $[d]=\{1,2,\ldots,d\}$. We use the notation $\|\cdot\|_p$ where $p\in\{0,1,2,\infty\}$, to denote the p-norm in $\mathbb{R}^d$. For any matrix $\mathbf{X}\in\mathbb{R}^{m\times n}$, let $\|\mathbf{X}\|_\infty=\max_{i\in[m]}\sum_{j=1}^n|X_{i,j}|$ be the infinity norm. Let $\{(\mathbf{x}_t,y_t)\}_{t\in[T]}$ be a sequence of examples, where $\mathbf{x}_t\in\mathcal{X}=\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}\|_\infty\leq 1\}$ is an instance. Let $\mathbb{I}_E$ be the indictor function. If the event E occurs, then $\mathbb{I}_E=1$. Otherwise, $\mathbb{I}_E=0$. For any subset $S\subset[d]$, let $S^c=[d]\setminus S$. For any $\mathbf{x}\in\mathbb{R}^d$ and $S\subset[d]$, we define $\mathbf{x}(S):=(x_1\cdot\mathbb{I}_{1\in S},x_2\cdot\mathbb{I}_{2\in S},\ldots,x_d\cdot\mathbb{I}_{d\in S})$. Let $\mathcal{N}(0,\sigma^2)$ be a

normal distribution with variance σ^2, $\mathbf{a}_d$ be the d-dimensional vector of a's, and $\mathbf{0}_{m\times n}$ be a matrix whose elements are all zeros. Let $\mathrm{Supp}(\mathbf{w}) = \{i \in [d] : w_i \neq 0\}$ be the support set of a vector $\mathbf{w}$.

2.2. Sufficient Conditions for Computationally Efficient Algorithms for OSLR

Assumption 1 (Realizable) *For each $(\mathbf{x}_t, y_t)$, $t = 1, 2, ..., T$, there exists a $\mathbf{w}^*$ such that*

$$y_t = \langle \mathbf{w}^*, \mathbf{x}_t \rangle + \eta_t,$$

in which $\eta_1, \eta_2, \ldots, \eta_T$ are independent noises from $\mathcal{N}(0, \sigma^2)$, $\|\mathbf{x}_t\|_\infty \leq 1$ and $\|\mathbf{w}^*\|_1 \leq 1$.

The normal distribution implies, with probability at least $1 - \delta$, $|y_t| \leq 1 + \sigma\sqrt{2 \ln \frac{1}{\delta}}$ for a fixed t.

Assumption 2 (Sparsity) *$\mathbf{w}^*$ is k-sparse, i.e., $\|\mathbf{w}^*\|_0 \leq k$.*

Assumption 3 ((δ_S, S, α)-compatibility condition (van de Geer and Bühlmann, 2009)) *Let $S \subset [d]$ be an arbitrary subset and $\Omega_S = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} \neq \mathbf{0}_d, \|\mathbf{w}(S^c)\|_1 \leq \alpha \|\mathbf{w}(S)\|_1\}$. The data matrix $\mathbf{X} \in \mathbb{R}^{d\times T}$ satisfies the (δ_S, S, α)-compatibility condition, if there is a constant $\delta_S > 0$ such that*

$$\min_{\mathbf{w}\in\Omega_S} \frac{|S| \cdot \|\mathbf{X}^\top \mathbf{w}\|_2^2}{T \cdot \|\mathbf{w}(S)\|_1^2} = \delta_S^2.$$

The (δ_S, S, α)-compatibility condition requires that the covariance matrix $\mathbf{X}\mathbf{X}^\top$ enjoys a kind of "restricted" positive definiteness on the restricted set Ω_S, and the ℓ_1-norm and the norm defined by $\mathbf{X}\mathbf{X}^\top$ to be compatible. As the equivalence of ℓ_1-norm and ℓ_2-norm of a vector, the compatibility condition is essentially equivalent to the restricted eigenvalues condition. For the sake of fair comparison with the results by (Ito et al., 2017), we adopt the compatibility condition. We can obtain similar or better results under the restricted eigenvalues condition by slightly adjusting the analysis. The compatibility condition is weaker than the three popular assumptions adopted for solving OSLR, including the linear independence of features, RIP, and RSC.

In this work, we only consider the case $3 \leq k \leq d - 3$. For $k \leq 2$ and $k \geq d - 2$, it is easy to establish an algorithm running in time $O(d)$ or $O(d^2)$. Specifically, we reduce OSLR to an instance of the adversarial multi-armed bandits problem (Auer et al., 2002) by enumerating all of the $N = \binom{d}{k}$ combinations of features and corresponding each combination as an arm. At each round $t \in [T]$, we maintain a parameter vector $\mathbf{w}_{t,i}$ for each combination $i \in [N]$. First, we select a $\mathbf{w}_{t,I_t}$, $I_t \in [N]$, and make a prediction $\hat{y}_t$. After observing y_t, $\mathbf{w}_{t,I_t}$ can be updated by online gradient descent (Zinkevich, 2003). The technical challenge is that $\mathbf{w}_{t,i}$ can not be updated for all $i \neq I_t$. We must carefully design a master algorithm for choosing I_t and an elaborate parameter updating strategy. The same technical challenge exists in the problem of corralling a band of bandit algorithms (Agarwal et al., 2017; Foster et al., 2020; Luo et al., 2022). Luckily, by the master algorithm proposed in (Foster et al., 2020), it is possible to obtain an expected regret of $O\left(\sqrt{NT}\right)$. A previous algorithm (Ito et al., 2018) uses a batching technique to combine any bandit algorithms and online learning algorithms, like ONS, but only provides a regret of $\tilde{O}\left(N^{\frac{1}{3}}T^{\frac{2}{3}}\right)$.

3. Algorithm

For clarity, we first provide a comprehensive overview of the key components and workflow of the proposed algorithm.

3.1. Overview of Algorithm

We first introduce some notations. Let $\mathcal{T} = \left\{s^2 : s = 1, \ldots, \lfloor\sqrt{T}\rfloor\right\}$ and $\mathcal{I}_s = \left\{1, 2^2, \ldots, s^2\right\}$. For any $s \in \left\{1, \ldots, \lfloor\sqrt{T}\rfloor\right\}$, let $\mathcal{T}_s = \left\{s^2+1, s^2+2, \ldots, (s+1)^2-1\right\}$. If $s = \lfloor\sqrt{T}\rfloor$, then we define $(s+1)^2 - 1 := T$. It is obvious that $[T] = \mathcal{T} \cup_{s=1}^{\lfloor\sqrt{T}\rfloor} \mathcal{T}_s$.

Overall, our algorithm alternately executes an exploration round and an exploitation epoch. Since the algorithms only access k attributes of each instance, it is impossible to conduct an exploration process per-round. Under certain assumptions, both LASSO (Tibshirani, 1996; Bunea et al., 2007) and the Dantzig Selector (Candès and Tao, 2007) can learn an estimator denoted by $\hat{\mathbf{w}}$ satisfying $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_p = O\left(1/\sqrt{T}\right)$ using T examples where $p \in \{1, 2\}$. Therefore, it is enough to conduct an exploration process across several rounds. Specifically, our algorithm will execute exploration at each round $t \in \mathcal{T}$. At any round $t \in \mathcal{T}$, our algorithm will sample k attributes and solve the Dantzig Selector, i.e., (5), to obtain an estimator of $\mathbf{w}^*$ denoted by $\hat{\mathbf{w}}_s$ where $s = \sqrt{t}$. The key technical contribution is the development of a novel algorithm-dependent sampling scheme.

Given $\hat{\mathbf{w}}_s$, we further construct an estimator of the true support set S, denoted by S_s. Then our algorithm transitions to the exploitation epoch $\mathcal{T}_s$ where our algorithm always observes the attributes indexed by S_s, and runs ONS initialized with $\hat{\mathbf{w}}_s(S_s)$. Although our algorithm uses ONS to learn $\mathbf{w}^*$, we still term the epoch "exploitation" because it fully relies on S_s and $\hat{\mathbf{w}}_s(S_s)$. We will propose a careful initialization scheme for ONS when our algorithm enters the next exploitation epoch $\mathcal{T}_{s+1}$.

However, solving (5) requires the prior of $\mathbf{w}^*$. To be specific, there is a parameter γ_s depending on $\|\Delta_\tau(S)\|_1, \tau < s$, where $\Delta_\tau(S) := \hat{\mathbf{w}}_\tau(S) - \mathbf{w}^*$. To tune γ_s adaptively, we first assume that $\|\Delta_\tau(S)\|_1, \tau < s$, are known. Then we can prove that $\|\Delta_s(S)\|_1$ satisfies the following inequality

$$\begin{aligned}\frac{\delta_S^2\|\Delta_s(S)\|_1}{k} \le c_1\frac{(d-1)(d-2)}{(k-1)(k-2)s}\ln\frac{d}{\delta} + c_2\sqrt{\frac{d-1}{s(k-1)}\ln\frac{d}{\delta}} + \sqrt{\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}} + \\ \frac{c_3}{s}\sqrt{\frac{(d-1)(d-2)}{(k-1)(k-2)}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}} + c_4\sqrt{\frac{(d-1)(d-2)}{(k-1)(k-2)s}\ln\frac{d^2}{\delta}}\|\Delta_s(S)\|_1.\end{aligned} \tag{3}$$

We can solve (3) by an induction method, yielding an ideal convergence rate on $\|\Delta_s(S)\|_1$ that nearly matches the final convergence rate in Lemma 2. We term it "an ideal convergence rate" as it requires the prior of $\mathbf{w}^*$. Then we use the ideal convergence rate to estimate $\|\Delta_\tau(S)\|_1$ and γ_s, which only depend on $\tau < s$ and some known constants. For simplicity, we will not give the ideal convergence rate, but implicitly use it to define the estimator of γ_s in Section 3.4.

3.2. Exploration

We use the Dantzig Selector to learn a sequence of estimators, as it can obtain a smaller constant factor in the convergence rate compared to LASSO. For the sake of clarity, we first introduce the Dantzig Selector in the full information setting. For a fixed $t \in \mathcal{T}$, let $\mathbf{X}_{\mathcal{I}_s} = (\mathbf{x}_1, \mathbf{x}_{2^2}, \ldots, \mathbf{x}_{s^2}) \in \mathbb{R}^{d\times s}$ and $\mathbf{Y}_{\mathcal{I}_s} = (y_1, y_{2^2}, \ldots, y_{s^2})^\top \in \mathbb{R}^s$, in which $s = \sqrt{t}$. The Dantzig Selector can return an estimator of $\mathbf{w}^*$, denoted by $\hat{\mathbf{w}}_s$, by solving the following constrained problem.

$$\min_{\mathbf{w}\in\mathbb{R}^d}\|\mathbf{w}\|_1, \quad \text{s.t.} \left\|\frac{1}{s}\mathbf{X}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s} - \frac{1}{s}\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top\mathbf{w}\right\|_\infty \le \gamma_s.$$

Unluckily, the algorithms for OSLR can not directly observe $\mathbf{X}_{\mathcal{I}_s}$. Therefore, it is necessary to construct estimators of $\mathbf{X}_{\mathcal{I}_s}$ and $\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top$. Previous algorithms (Ito et al., 2017; Kale et al., 2017; Murata and Suzuki, 2018) uniformly sample k attributes from $\mathbf{x}_{\tau^2}$ for constructing unbiased estimators of $\mathbf{x}_{\tau^2}$ and $\mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top$ for all $\tau = 1, 2, \ldots, s$. Nevertheless, employing such a basic sampling scheme degrades the convergence rate concerning its reliance on the dimension d. To address this issue, we will define an algorithm-dependent sampling scheme. A better sampling probability should be defined as a function of $\mathbf{w}^*$, such as $\mathbf{q}^* = (q_1^*, \ldots, q_d^*)$ in which $q_i^* = \frac{|w_i^*|}{\|\mathbf{w}^*\|_1}$, $i \in [d]$, implying $x_{\tau^2,i}$ will not be selected for any $i \notin S$. However, it is infeasible to construct $\mathbf{q}^*$ as $\mathbf{w}^*$ is unknown. Benefit from the sequential nature of online learning, we can utilize the estimator in the last round, denoted by $\hat{\mathbf{w}}_{s-1}$, as a reliable approximation of $\mathbf{w}^*$ and construct a sampling distribution [2], denoted by $\mathbf{q}_s = (q_{s,1}, \ldots, q_{s,d})$, in which $q_{s,i} = \frac{|\hat{w}_{s-1,i}|}{\|\hat{\mathbf{w}}_{s-1}\|_1}$. It is still far from optimality to sample from $\mathbf{q}_s$ due to $\hat{\mathbf{w}}_{s-1} \neq \mathbf{w}^*$. The error between $\hat{w}_{s-1,i}$ and w_i^*, $i \in [d]$, makes the estimators suffering large variances. Next we explain our sampling scheme.

At any round $t \in \mathcal{T}$, we first sample a feature x_{t,I_1} from $\{x_{t,1}, \ldots, x_{t,d}\}$ following $\mathbf{q}_s$, and then uniformly sample $k-1$ features denoted by $x_{t,I_2}, \ldots, x_{t,I_k}$ from $\{x_{t,1}, \ldots, x_{t,d}\} \setminus \{x_{t,I_1}\}$ without replacement, in which $s = \sqrt{t}$. For simplicity, let $B_s = \{I_1, \ldots, I_k\}$, $\mathbb{P}[i \in B_s]$ be probability that the i-th feature is observed and $\mathbb{P}[i, j \in B_s]$ be the probability that both the i-th feature and the j-th feature are observed. Initializing $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$. Next we construct unbiased estimators of $\mathbf{x}_t$ and $\mathbf{x}_t\mathbf{x}_t^\top$. Let $\hat{\mathbf{x}}_t = (\hat{x}_{t,1}, \hat{x}_{t,2}, \ldots, \hat{x}_{t,d})^\top \in \mathbb{R}^d$ and $\mathbf{h}_t \in \mathbb{R}^{d\times d}$ satisfy

$$\begin{aligned} \forall i \in [d], \qquad \hat{x}_{t,i} =& \frac{x_{t,i}}{\mathbb{P}[i \in B_s]} \cdot \mathbb{I}_{i\in B_s}, \quad h_t[i,i] = \frac{x_{t,i}^2}{\mathbb{P}[i \in B_s]} \cdot \mathbb{I}_{i\in B_s}, \\ \forall i \neq j \in [d], \quad h_t[i,j] =& \frac{x_{t,i}x_{t,j}}{\mathbb{P}[i,j \in B_s]} \cdot \mathbb{I}_{i,j\in B_s}. \end{aligned} \tag{4}$$

Then we define unbiased estimators of $\mathbf{X}_{\mathcal{I}_s}$ and $\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top$ as follows,

$$\hat{\mathbf{X}}_{\mathcal{I}_s} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_{2^2}, \hat{\mathbf{x}}_{3^2} \ldots, \hat{\mathbf{x}}_{s^2}] \in \mathbb{R}^{d\times s}, \quad \mathbf{H}_{\mathcal{I}_s} = \sum_{\tau=1}^{s} \mathbf{h}_{\tau^2} \in \mathbb{R}^{d\times d}.$$

Let $g_{d,k} = \frac{(d-1)(d-2)}{(k-1)(k-2)}$. Now we define the following time-variant Dantzig Selector ($\mathrm{DS}(\gamma_s)$)

$$\mathrm{DS}(\gamma_s): \min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{w}\|_1 \quad \text{s.t.} \quad \left\| \frac{1}{s}\hat{\mathbf{X}}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s} - \frac{1}{s}\mathbf{H}_{\mathcal{I}_s}\mathbf{w} \right\|_\infty \leq \gamma_s, \tag{5}$$

where γ_s is a time-variant threshold defined as follows

$$\gamma_s = \left(\frac{8}{3} + 2\sigma\right)\frac{g_{d,k}}{s}\ln\frac{d}{\delta} + \frac{6.9 + 1.2\sigma}{\sqrt{s}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}} + \frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2 \ln\frac{d}{\delta}}, \tag{6}$$

in which $\Delta_{\tau-1}(S) = \hat{\mathbf{w}}_{\tau-1}(S) - \mathbf{w}^*$. Denoted by $\hat{\mathbf{w}}_s$ the optimal solution of $\mathrm{DS}(\gamma_s)$. $\hat{\mathbf{X}}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s}$ can be computed incrementally in a space and per-round time complexity of $O(d)$. Besides, computing

2. A similar but fundamentally different idea was adopted by the OLin-LASSO Algorithm (Yang et al., 2023), in which $\hat{\mathbf{w}}_{s-1}$ was used to approximate the square loss function at round s.

$\mathbf{H}_{\mathcal{I}_s}$ necessitates a space and per-round time complexity of $O(d^2)$. $\mathrm{DS}(\gamma_s)$ can be recast as a linear programming (Candès and Tao, 2007), and be solved in polynomial time. For instance, the interior point method (Karmarkar, 1984) requires time in $O(d^{3.5}L)$ and the algorithm by Vaidya (1989) requires time in $O(d^{2.5}L)$, in which L is the number of bits in the input of the linear programming.

3.3. Exploitation

Let $S_0 \subseteq [d]$ be an arbitrary subset satisfying $|S_0| = k$. For $s \geq 1$, it can be proved that $\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1 \to 0$ in probability, suggesting a natural estimation of the true support set S, denoted by S_s,

$$S_s \subseteq [d], \quad \text{s.t. } |S_s| = k, \ \forall i \in S_s, j \in [d] \setminus S_s, \quad |\hat{w}_{s,i}| \geq |\hat{w}_{s,j}| \,.$$

Before S_s converges to S, we can trust $\hat{\mathbf{w}}_s(S_s)$ throughout the s-th epoch $\mathcal{T}_s$, that is, $\hat{\mathbf{w}}_s(S_s)$ can serve as a good predictor for all $\mathbf{x}_t$, $t \in \mathcal{T}_s$. The regret in $\mathcal{T}_s$ obviously depends on the convergence rate $\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1$. There is a s_2, such that $S_s = S$ for any $s > s_2$ with a high probability. In this case, $\hat{\mathbf{w}}_s(S_s)$ is not the best predictor for all $\mathbf{x}_t$, since $\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1$ decays with s, not t (the number of observed examples). To address this issue, we can solve an ordinary online linear regression on the observations $\{(\mathbf{x}_t(S_s), y_t)\}_{t \in \mathcal{T}_s}$, and use an online learning algorithm to learn $\mathbf{w}^*$. Since the square loss function is exp-concave, we will use ONS [3]. The per-round time complexity is $O(k^2)$.

As s_2 is unknown, we will update $\hat{\mathbf{w}}_s(S_s)$ during the epoch $\mathcal{T}_s$ for all $s \geq 1$. We first define a feasible set $\mathcal{W}_s$ as follows

$$\forall s \geq 1, \quad \mathcal{W}_s = \left\{ \mathbf{w} \in \mathbb{R}^d : \mathrm{Supp}(\mathbf{w}) \subset S_s, \forall t \in \mathcal{T}_s, \langle \mathbf{w}, \mathbf{x}_t(S_s) \rangle \leq 1 \right\}.$$

It is obvious that $\hat{\mathbf{w}}_s(S_s) \in \mathcal{W}_s$. Initializing a parameter vector $\bar{\mathbf{w}}_{s^2+1}(S_s) = \hat{\mathbf{w}}_s(S_s)$, and a covariance matrix $\mathbf{A}_s = \varepsilon \cdot \mathbf{I}_{k \times k}$. At each round $t \in \mathcal{T}_s$, the prediction is given by $\hat{y}_t = \langle \bar{\mathbf{w}}_t(S_s), \mathbf{x}_t \rangle$. Let $\mathbf{g}_t = 2(\hat{y}_t - y_t)\mathbf{x}_t(S_s)$ be the gradient. The parameters are updated as follows,

$$\mathbf{A}_t = \mathbf{A}_{t-1} + \rho \mathbf{g}_t \mathbf{g}_t^\top, \ \bar{\mathbf{w}}_{t+1}(S_s) = \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{W}_s} \left\| \mathbf{w} - \bar{\mathbf{w}}_t(S_s) + \mathbf{A}_t^{-1}\mathbf{g}_t \right\|_{\mathbf{A}_t} := \mathcal{P}_s^t(\bar{\mathbf{w}}_t(S_s) - \mathbf{A}_t^{-1}\mathbf{g}_t),$$

in which $\mathcal{P}_s^t(\cdot)$ is a projector operator (Luo et al., 2016) defined as follows,

$$\mathcal{P}_s^t(\mathbf{w}) = \mathbf{w} - \frac{\tau(\langle \mathbf{w}, \mathbf{x}_t(S_s) \rangle)}{\langle \mathbf{x}_t(S_s), \mathbf{A}_t^{-1}\mathbf{x}_t(S_s) \rangle} \mathbf{A}_t^{-1}\mathbf{x}_t(S_s), \quad \tau(y) = \mathrm{sign}(y) \cdot \max\{|u| - 1, 0\}.$$

If $S_s = S_{s-1}$, then resetting $\bar{\mathbf{w}}_{s^2+1} = \hat{\mathbf{w}}_s(S_s)$ and $\mathbf{A}_{s^2} = \varepsilon \cdot \mathbf{I}_{k \times k}$ must increase the regret. To address this issue, we redefine the initial configurations as follows

$$\begin{cases} \bar{\mathbf{w}}_{s^2+1}(S_s) = \hat{\mathbf{w}}_s(S_s), & \mathbf{A}_{s^2} = \varepsilon \cdot \mathbf{I}_{k \times k}, \quad \text{if } S_s \neq S_{s-1}, \\ \bar{\mathbf{w}}_{s^2+1}(S_s) = \bar{\mathbf{w}}_{s^2}(S_{s-1}), & \mathbf{A}_{s^2} = \mathbf{A}_{s^2-1}, \quad \text{otherwise.} \end{cases}$$

3. If we use online gradient descent (Zinkevich, 2003), then the algorithm only requires a per-round time complexity in $O(k)$, but incurs larger regret.

3.4. Adaptive Parameters-tuning

By (6), γ_s requires the prior information of $\mathbf{w}^*$. It is necessary to construct an estimator of γ_s, denoted by $\hat{\gamma}_s$. By theoretical analyses, we must ensure that $\mathrm{DS}(\hat{\gamma}_s)$ has a solution $\hat{\mathbf{w}}_s$ satisfying $\|\hat{\mathbf{w}}_s\|_1 \leq \|\mathbf{w}^*\|_1$. To this end, it is sufficient to make $\hat{\gamma}_s$ be a tight upper bound of γ_s. By the definition of γ_s, it is further reduced to provide a tight upper bound on $\frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2 \ln\frac{d}{\delta}}$. Let ν_s be an estimator such that $\nu_s \geq \frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2 \ln\frac{d}{\delta}}$. Then $\hat{\gamma}_s$ can be defined by

$$\hat{\gamma}_s = \left(\frac{8}{3}+2\sigma\right)\frac{g_{d,k}}{s}\ln\frac{d}{\delta} + \frac{6.9+1.2\sigma}{\sqrt{s}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}} + \nu_s. \tag{7}$$

The challenge is how to define ν_s. Note that if γ_s is known, we can obtain the ideal convergence rate of $\|\Delta_s(S)\|_1$. To be specific, we first solve (3) by an induction method, and obtain the ideal convergence rate. Then we use it to define ν_s. We will verify whether $\hat{\gamma}_s \geq \gamma_s$ for all $s \geq 1$, and concurrently establish the real convergence rate of $\|\Delta_s(S)\|_1$ by an induction method. The real convergence rate nearly matches the ideal one. Let $\delta \in (0,1)$, $\mu_1 = \frac{9}{9-2\sqrt{3}}$, and

$$\mu_2 = \frac{1}{1-\frac{\sqrt{6}}{9\sqrt{\frac{d-2}{k-2}\ln\frac{d^2}{\delta}}}}, \quad s_0 = \frac{24^2k^2g_{d,k}}{\delta_S^4}\ln\frac{d^2}{\delta}, \quad s_1 = \frac{24^2k^2g_{d,k}}{\delta_S^4}\frac{d-2}{k-2}\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}. \tag{8}$$

We consider the problems where the number of examples is sufficient large, i.e., $T > (s_1+1)^2$. Let

$$\begin{cases} a_1 = \left(\frac{64}{3}+\frac{32}{3}\sigma\right)\ln\frac{d}{\delta}, \quad a_2 = \frac{16\,(6.9+1.2\sigma)}{3}\sqrt{\ln\frac{d}{\delta}}, \quad a_3 = \frac{8}{3}\sqrt{3\ln\frac{d}{\delta}}, \\ a_4 = \delta_S^2\frac{a_1}{k} + 24a_2\sqrt{\frac{k-2}{d-2}\ln\frac{d^2}{\delta}} + 4a_3\left(24\sqrt{\ln\frac{d^2}{\delta}} + \frac{\delta_S^2}{k\sqrt{g_{d,k}}}\right), \\ a_5 = \frac{9}{9-2\sqrt{3}}\left(\delta_S^2\frac{8+4\sigma}{9k} + \frac{32}{\sqrt{3}} + \frac{4\sqrt{3}\delta_S^2}{9k\sqrt{g_{d,k}\ln\frac{d^2}{\delta}}}\right) + a_2 + \frac{2\sqrt{3}a_2}{9-2\sqrt{3}}\sqrt{\frac{k-2}{(d-2)\ln\frac{d^2}{\delta}}}. \end{cases} \tag{9}$$

Then ν_s is defined as follows

$$\nu_s = \begin{cases} \frac{2}{\sqrt{s}}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}, & s \leq [1, s_0], \\ \frac{1}{s}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d,k}\ln\frac{d^2}{\delta}}+2\right), & s = s_0+1, \\ \frac{s_0+1}{s}\nu_{s_0+1} + \frac{1}{s}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}\cdot\frac{\mu_1 a_4}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s-1}\frac{k^4g_{d,k}^2}{\tau^2}}, & s \in (s_0+1, s_1], \\ \frac{1}{s}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d,k}\ln\frac{d^2}{\delta}}+2+\frac{\mu_1 a_4 k^2 g_{d,k}}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s_1}\frac{1}{\tau^2}}\right), & s = s_1+1, \\ \frac{s_1+1}{s}\nu_{s_1+1} + \frac{1}{s}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}\cdot\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\sum_{\tau=s_1+1}^{s-1}\frac{k^2(d-1)}{\tau(k-1)}}, & s > s_1+1. \end{cases}$$

Now we can solve $\mathrm{DS}(\hat{\gamma}_s)$ and obtain the solution $\hat{\mathbf{w}}_s$. We name this algorithm DS-OSLRC (Dantzig Selector for OSLR with Compatibility condition) and show the pseudo-code in Algorithm 1.

Algorithm 1: DS-OSLRC

Input: k, d, σ, δ_S, ρ, δ

Initialize $\mathbf{A}_1 = \varepsilon \cdot \mathbf{I}_{k\times k}$, $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$, $\hat{\mathbf{x}}_{\mathcal{I}_0} = \mathbf{0}_d$, $\mathbf{H}_{\mathcal{I}_0} = \mathbf{0}_{d\times d}$, S_0;
for $t = 1, 2, \ldots, T$ **do**
  if $t \in \mathcal{T}$ **then**
    $s = \sqrt{t}$;
    Obtain $B_s \subseteq [d]$ from SAMPLING($k, d, \hat{\mathbf{w}}_{s-1}$);
    Output the prediction $\hat{y}_t = \langle \hat{\mathbf{w}}_{s-1}(B_s), \mathbf{x}_t(B_s)\rangle$;
    Compute $\hat{\mathbf{X}}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s} = \hat{\mathbf{X}}_{\mathcal{I}_{s-1}}\mathbf{Y}_{\mathcal{I}_{s-1}} + \hat{\mathbf{x}}_{s^2}y_{s^2}$ where $\hat{\mathbf{x}}_{s^2}$ is computed by combining (4) with (10);
    Compute $\mathbf{H}_{\mathcal{I}_s} = \mathbf{H}_{\mathcal{I}_{s-1}} + \mathbf{h}_{s^2}$ where $\mathbf{h}_{s^2}$ is computed by combining (4) with (10);
    Compute $\hat{\gamma}_s$ by (7);
    Obtain $\hat{\mathbf{w}}_s$ by solving DS($\hat{\gamma}_s$) and select S_s;
    if $S_s \neq S_{s-1}$ **then**
      Initialize $\bar{\mathbf{w}}_{s^2+1}(S_s) = \hat{\mathbf{w}}_s(S_s)$;
      Initialize $\mathbf{A}_{s^2} = \varepsilon \cdot \mathbf{I}_{k\times k}$;
    end
    else
      Initialize $\bar{\mathbf{w}}_{s^2+1}(S_s) = \bar{\mathbf{w}}_{s^2}(S_{s-1})$;
      Initialize $\mathbf{A}_{s^2} = \mathbf{A}_{s^2-1}$;
    end
  end
  else
    Output the prediction $\hat{y}_t = \langle \bar{\mathbf{w}}_t(S_s), \mathbf{x}_t(S_s)\rangle$;
    Compute $\mathbf{g}_t = 2(\hat{y}_t - y_t)\mathbf{x}_t(S_s)$;
    Update $\mathbf{A}_t = \mathbf{A}_{t-1} + \rho \cdot \mathbf{g}_t\mathbf{g}_t^\top$;
    Compute $\bar{\mathbf{w}}_{t+1}(S_s) = \mathcal{P}_s^t\left(\bar{\mathbf{w}}_t(S_s) - \mathbf{A}_t^{-1}\mathbf{g}_t\right)$;
  end
end

Algorithm 2: SAMPLING

Input: k, d, $\mathbf{w}$

Initialize $B = \emptyset$;
Compute $\mathbf{q} = \frac{\mathbf{w}}{\|\mathbf{w}\|_1}$;
Sample $I_t \in [d]$ following $\mathbf{q}$;
Update $B = B \cup \{I_t\}$;
for $r = 1, \ldots, k-1$ **do**
  Sampling $I_r \in [d] \setminus B$ uniformly;
  $B = B \cup \{I_{r+1}\}$;
end
Return B;

4. Main Results

In this section, we give the ℓ_1-norm error of $\hat{\mathbf{w}}_s(S)$ and the regret bound of DS-OSLRC.

4.1. ℓ_1-norm Error Bound

Lemma 1 gives the sampling probabilities, making it possible to compute $\hat{\mathbf{x}}_{s^2}$ and $\mathbf{h}_{s^2}$. The analysis is non-trivial, given that DS-OSLRC samples attributes without replacement. We do not prove Lemma 1, but instead, prove a more general version, namely, Lemma 9 in Appendix B.

Lemma 1 *For any $s \geq 1$, let B_s be the output of Algorithm 2.*

$$\begin{aligned} \forall i \in [d], \quad \mathbb{P}[i \in B_s] =& \frac{d-k}{d-1} q_{s,i} + \frac{k-1}{d-1}, \\ \forall i \neq j \in [d], \quad \mathbb{P}[i,j \in B_s, i \neq j] =& \frac{(k-1)(k-2)}{(d-1)(d-2)} + \frac{(k-1)(d-k)}{(d-1)(d-2)} \cdot (q_{s,i} + q_{s,j}). \end{aligned} \tag{10}$$

Next we provide the ℓ_1-norm error of $\hat{\mathbf{w}}_s(S)$, which serves as the foundation for regret analysis.

Lemma 2 (Estimation Error) *Let $S = \mathrm{Supp}(\mathbf{w}^*)$, $0 < \delta < 1$ and $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$. If $3 \leq k \leq d-3$, $T > (s_1+1)^2$, Assumptions 1-2 hold, and $\mathbf{X}_{\mathcal{I}_s}$ satisfies the $(\delta_S, S, 1)$-compatibility condition for all $s \geq 1$, then with probability at least $1 - \sqrt{T}(6 + \log_{1.5}\frac{d+k}{k-2})\delta$, DS-OSLRC guarantees*

$$\forall s \geq 1,\ \|\Delta_s(S)\|_1 \leq \begin{cases} \frac{16+8\sigma}{\delta_S^2}\frac{k g_{d,k}}{s}\ln\frac{d}{\delta} + \frac{(26+4.8\sigma)k}{\delta_S^2}\sqrt{\frac{(d-1)\ln\frac{d}{\delta}}{s(k-1)}} + \frac{22k}{\delta_S^2}\sqrt{\frac{g_{d,k}}{s}\ln\frac{d}{\delta}}, & s \leq [1, s_0] \\ \frac{9}{9-2\sqrt{3}}\frac{a_4}{\delta_S^4}\frac{k^2 g_{d,k}}{s}, & s \in (s_0, s_1] \\ \mu_2 \cdot \frac{a_5}{\delta_S^2}\sqrt{\frac{k^2(d-1)}{s(k-1)}}, & s > s_1, \end{cases}$$

in which μ_2, s_0 and s_1 follow (8), a_4 and a_5 follow (9).

By Lemma 2, it is easy to establish the ℓ_1-norm error bound of $\hat{\mathbf{w}}_s$. To be specific, by Lemma 16, we have $\|\Delta_s\|_1 \leq 2\|\Delta_s(S)\|_1$. We can also obtain the ℓ_2-norm error bound by the inequality $\|\Delta_s\|_2 \leq \|\Delta_s\|_1$. It is worth mentioning that by the restricted eigenvalues condition (Bickel et al., 2009), we can obtain a tighter ℓ_2-norm error bound.

The Algorithm 3 in (Ito et al., 2017) attains $\|\Delta_s\|_1 = O(s^{-\frac{1}{4}})$ (please refer to Lemma 14 in original paper), while our convergence rate is $\tilde{O}(s^{-\frac{1}{2}})$. The algorithm in (Kale et al., 2017) attains $\|\Delta_s\|_1 = O(\frac{d}{k_0}\sqrt{\frac{d}{k_0}\frac{k^2}{s}\ln\frac{d}{\delta}})$. For $s > s_1$, DS-OSLRC improves the convergence rate by a factor of $O(d)$. In the full information setting where algorithms can observe $\mathbf{x}_s$ for all $s = 1, \ldots, T$, the OLin-LASSO algorithm (Yang et al., 2023) attains $\|\Delta_s\|_1 = \tilde{O}(\sqrt{k/s})$. Our convergence rate only deteriorates by a factor of $O(\sqrt{d})$.

4.2. Regret Bounds of DS-OSLRC

Theorem 1 (Regret Bound w.r.t. $\mathbf{w}^*$) *Let $\varepsilon = k$, $\delta \in (0,1)$ and*

$$Y_\delta = 1 + \sigma\sqrt{2\ln\frac{1}{\delta}}, \quad \rho = \frac{1}{2(1+Y_\delta)^2}, \quad s_2 = 4\frac{(\mu_2 a_5)^2}{\delta_S^4}\frac{k^2(d-1)}{\min_{i\in S}|w_i^*|^2(k-1)}.$$

Under the same assumptions in Lemma 2 and the condition

$$T > (s_2+1)^2 > (s_1+1)^2,$$

with probability at least $1 - (T + \sqrt{T}(6 + \log_{1.5}\frac{d+k}{k-2}) + 1)\delta$, the regret of DS-OSLRC satisfies

$$\begin{aligned} \mathrm{Reg}(\mathbf{w}^*) \leq& 4\sqrt{T} + \frac{2(\mu_2 a_5)^2}{\delta_S^4} \cdot \frac{k^3(d-1)\ln\left(4(1+Y_\delta)^2 T + 1\right)}{(k-1)\min_{i\in S}|w_i^*|^2} + \\ & \frac{22a_4^2}{\delta_S^8}k^4 g_{d,k}^2 \ln\frac{(d-2)\ln\frac{d}{\delta}}{k-2} + \frac{2(2\mu_2 a_5)^4}{\delta_S^8} \cdot \frac{k^4(d-1)^2}{\min_{i\in S}|w_i^*|^2(k-1)^2} + O(1), \end{aligned}$$

where μ_2 follows (8), a_4 and a_5 follow (9) and $O(1)$ hides the lower order constant terms.

For the sake of simplicity, we only consider the scenario where T is adequately large. It is easy to analyze the regret in the cases where $T \leq (s_2+1)^2$, or $s_2 < s_1$. If $t \in \mathcal{T}$, DS-OSLRC solves a linear programming. The time complexity is denoted by $O(\mathrm{LP}_d) = O(\mathrm{poly}(d))$. Otherwise, the per-round time complexity is $O(k^2)$. Thus the average per-round time complexity is only $O(k^2 + \frac{\mathrm{LP}_d}{\sqrt{T}})$. Let $\delta = \Theta(\frac{1}{T})$. Then DS-OSLRC achieves a $O\left(\sqrt{T} + \frac{k^2 d^2}{\delta_S^8 h(\mathbf{w}^*)^2} \ln^2 \frac{dT}{\delta}\right)$ regret bound within a time complexity of $O(\mathrm{poly}(d))$ per-iteration, making OSLR tractable. Note that if $k \ll d$ and $\sigma \leq 1$, then the numerical factor on the dominated term is $2(2\mu_2 a_5)^4 \ln^{-2} \frac{dT}{\delta} \approx 2(2a_2)^4 \ln^{-2} \frac{dT}{\delta} \approx 102^4$.

Ito et al. (2017) proposed two algorithms denoted by **alg2** and **alg3**, for OSLR, both of which enjoy an expected regret of $O(\sqrt{kT} + \mathrm{poly}(d, k))$ in time $O(d)$ per-iteration. Under the assumptions including (i) realizable, (ii) $\mathbf{x}_t \sim \mathcal{D}_{\mathbf{x}}$ for all $t \in [T]$, (iii) the features in $\mathbf{x}_t$ are linearly independent for all $t \in [T]$, and (iv) bounded noises, the expected regret of **alg2** satisfies

$$\mathbb{E}[\mathrm{Reg}(\mathbf{w}^*)] \leq 8\sqrt{kT} + \sum_{i \in S} \frac{8192^2 d^4 (d-1)^4}{\sigma_d^8 |w_i^*|^7 k^4 (k-1)^4} + O(1) = O\left(\sqrt{kT} + \frac{d^8}{\sigma_d^8 k^7 h(\mathbf{w}^*)^7}\right),$$

where σ_d^2 is the smallest eigenvalue of $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$. It must be $\sigma_d \leq \delta_S$. The linear independence of features condition is stronger than the compatibility condition. What's more, the regret bound is much worse than ours w.r.t. the dependence on d, $\min_{i \in S} |w_i^*|$ and T.

Under the same assumptions (except for the bounded noises and i.i.d. instances) with our algorithm, the expected regret of **alg3** satisfies

$$\mathbb{E}[\mathrm{Reg}(\mathbf{w}^*)] \leq 8\sqrt{kT} + \sum_{i \in S} \frac{128^2 \cdot 36^8 d^4 (d-1)^4}{\delta_S^8 |w_i^*|^7 (k-1)^4} + O(1) = O\left(\sqrt{kT} + \frac{d^8}{\delta_S^8 k^3 h(\mathbf{w}^*)^7}\right).$$

The regret bound is also much worse than ours w.r.t. the dependence on d, $\min_{i \in S} |w_i^*|$ and T. Besides, the constant factor is also much larger than ours.

Remark 1 *We can also analyze the regret w.r.t. the best k-sparse linear regressor. To be specific, by the regret analysis in (Kale et al., 2017), we obtain, with probability at least $1 - \delta$,*

$$\max_{\mathbf{w} \in \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_0 \leq k\}} \sum_{t=1}^{T} [\ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t) - \ell(\langle \mathbf{w}, \mathbf{x}_t \rangle, y_t)] \leq 2k\sigma^2 + 4k\sigma^2 \ln \frac{d}{\delta}. \tag{11}$$

By incorporating the upper bound on $\mathrm{Reg}(\mathbf{w}^*)$, i.e., Theorem 1, we can obtain the regret bound w.r.t. the best k-sparse linear regressor.

4.3. Discussion of Lower Bounds and Upper Bounds

Next we compare our upper bounds with lower bounds. The results are summarized in Table 3.

The lower bound on estimator error is for sparse linear regression (Candès and Davenport, 2013). OSLR is an online, partial information variant of sparse linear regression. Therefore, this lower bound is applicable to OSLR. To be specific, any algorithm for OSLR can return an estimator of $\mathbf{w}^*$, denoted by $\hat{\mathbf{w}}_T$, which can serve as a solution of sparse linear regression. Note that the original lower bound established in (Candès and Davenport, 2013) applies to any design matrix $\mathbf{X}$. We adapt the general bound to Gaussian designs whose entries are i.i.d. $\mathcal{N}(0, 1)$. As Gaussian

lower bound on estimation error	upper bound for OSLR	upper bound for (k, k_0, d)-OSLR
$\Omega\left(\sqrt{\frac{k}{T}\log\frac{d}{k}}\right)$ (Candès and Davenport, 2013)	$O\left(\sqrt{\frac{kd}{T}\log\frac{dT}{\delta}}\right)$	$O\left(\sqrt{\frac{k^2 d}{Tk_0}\log\frac{dT}{\delta}}\right)$
lower bound on regret	upper bound for OSLR	upper bound for (k, k_0, d)-OSLR
$\Omega\left(k\ln(T)\ln\frac{d}{k}\right)$	$O\left(\sqrt{T} + k^2 d^2 \ln^2\frac{dT}{\delta}\right)$	$O\left(\frac{k^2 d}{k_0}(\frac{d}{k_0} + \ln T)\ln\frac{dT}{\delta}\right)$

Table 3: Lower bounds and upper bounds on estimation error and regret. The estimation error is $\min_{\hat{\mathbf{w}}}\max_{\mathbf{w}^*}\mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2]$.

designs satisfy the compatibility condition, the lower bound remains applicable to OSLR under the assumptions in our paper. There is a gap of $O\left(\sqrt{d}\right)$ between the lower and upper bounds. We conjecture that the lower bound can be refined, given the partial information constraint inherent to OSLR. Under the condition that all of the attributes per instance can be observed, the OLin-LASSO algorithm (Yang et al., 2023) indeed attains the lower bound.

The lower bound on the regret remains unestablished. We can easily obtain a lower bound on regret by the lower bound on estimation error. Let $\mathbf{X}$ be a Gaussian design. Consider any algorithm that maintains an estimator $\hat{\mathbf{w}}_s$, and produces a prediction using $\hat{\mathbf{w}}_s(S_s)$ where $S_s \subseteq [d]$ and $|S_s| \leq k$ at each round $s \geq 1$. The expected regret satisfies

$$\mathbb{E}[\mathrm{Reg}(\mathbf{w}^*)] = \sum_{s=1}^{T}\mathbb{E}[\langle\hat{\mathbf{w}}_s(S_s) - \mathbf{w}^*, \mathbf{x}_s\rangle^2] = \Omega\left(\sum_{s=1}^{T}\frac{1}{s}k\ln\frac{d}{k}\right) = \Omega\left(k\ln(T+1)\ln\frac{d}{k}\right),$$

where the expectation is w.r.t. the noises and $\mathbf{X}$.

5. Conclusion and Future Work

In this paper, we have proposed a new polynomial-time algorithm for OSLR that significantly improves previous regret bounds in terms of both problem-dependent parameters and constant factors under some mild assumptions. Notably, we assume the data matrix satisfies the compatibility condition that is less restrictive than the linear independence of features condition and RIP utilized in prior work. We further extend the algorithm to (k, k_0, d)-OSLR and improve previous regret bounds.

Our work opens several directions for future research. The first one is to develop more efficient algorithms for OSLR that can avoid solving a linear programming while maintaining the regret bound under the same assumptions. The second one is to establish more efficient algorithms for (k, k_0, d)-OSLR. Our algorithm requires solving a linear programming at each round (please refer to Appendix A). The third one is to establish tight lower bounds on both estimation error and regret.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants No. 62236003 and 62076181. We also appreciate all the anonymous reviewers for their valuable comments and constructive suggestions.

References

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E. Schapire. Corralling a band of bandit algorithms. In *Proceedings of the 30th Annual Conference on Learning Theory,*, pages 12–38, 2017.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Peter J. Bickel, Yaacov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Emmanuel J. Candès and Mark A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Emmanuel J. Candès and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

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. In *Proceedings of the 27th International Conference on Machine Learning*, pages 183–190, 2010.

Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 12:2857–2878, 2011.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Annals of Statistics*, 46(2):814–841, 2018.

Dean Foster, Satyen Kale, and Howard Karloff. Online sparse linear regression. In *Proceedings of the 29th Annual Conference on Learning Theory*, pages 960–970, 2016.

Dean P. Foster, Howard J. Karloff, and Justin Thaler. Variable selection is hard. In *Proceedings of The 28th Conference on Learning Theory*, pages 696–709, 2015.

Dylan J. Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

Elad Hazan and Tomer Koren. Linear regression with limited observation. In *Proceedings of the 29th International Conference on Machine Learning*, pages 807–814, 2012.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Akihiro Yabe, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Efficient sublinear-regret algorithms for online sparse linear regression with limited observation. *Advances in Neural Information Processing Systems*, 30:4099–4108, 2017.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Akihiro Yabe, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Online regression with partial information: Generalization and linear projection. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1599–1607, 2018.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 24.1–24.34, 2012.

Satyen Kale. Open problem: Efficient online sparse regression. In *Proceedings of the 27th Annual Conference on Learning Theory*, pages 1299–1301, 2014.

Satyen Kale, Zohar Karnin, Tengyuan Liang, and Dávid Pál. Adaptive feature selection: Computationally efficient online sparse linear regression under RIP. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1780–1788, 2017.

Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing*, pages 302–311, 1984.

Junfan Li, Zheshun Wu, Zenglin Xu, and Irwin King. On the necessity of collaboration for online model selection with decentralized data. *Advances in Neural Information Processing Systems*, 37:85583–85629, 2024.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. *Advances in Neural Information Processing Systems*, 29:902–910, 2016.

Haipeng Luo, Mengxiao Zhang, Peng Zhao, and Zhi-Hua Zhou. Corralling a larger band of bandits: A case study on switching regret for linear bandits. In *Proceedings of the 35th Annual Conference Computational Learning Theory*, pages 3635–3684, 2022.

Tomoya Murata and Taiji Suzuki. Sample efficient stochastic gradient iterative hard thresholding method for stochastic sparse linear regression with limited attribute observation. *Advances in Neural Information Processing Systems*, 31:5317–5326, 2018.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24 (2):227–234, 1995.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 57(1):267–288, 1996.

Pravin M. Vaidya. Speeding-up linear programming using fast matrix multiplication (extended abstract). In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 332–337, 1989.

Sara van de Geer. The deterministic lasso. *American Statistical Association*, 2007.

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Haiqin Yang, Zenglin Xu, Irwin King, and Michael R. Lyu. Online learning for group lasso. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1191–1198, 2010.

Shuoguang Yang, Yuhao Yan, Xiuneng Zhu, and Qiang Sun. Online linearized LASSO. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 7594–7610, 2023.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, 2003.

Navid Zolghadr, Gábor Bartók, Russell Greiner, András György, and Csaba Szepesvári. Online Learning with Costly Features and Labels. *Advances in Neural Information Processing Systems*, 26:1241–1249, 2013.

Appendix A. Extension to OSLR with Additional Observations

In this section, we extend DS-OSLRC to (k, k_0, d)-OSLR (Foster et al., 2016) which is also called proper online sparse linear regression (POSLR) (Kale et al., 2017). At each round $s \geq 1$, the learner chooses k attributes from $\mathbf{x}_s$ and makes a prediction $\hat{y}_s$. Then the adversary gives the true output y_s. After that the learner can observe another $k_0 > 1$ attributes. In this work, we consider the case $k_0 \geq 3$ and $k_0 = O(k \ln d)$. We will propose a new algorithm for (k, k_0, d)-OSLR, which can improve the previous regret bounds under weaker assumptions.

Let $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$, and $S_0 \subseteq [d]$ be an arbitrary subset satisfying $|S_0| = k$. For each $s \geq 1$, let $S_s \subseteq [d]$ follow the definition in DS-OSLRC. At the beginning of the s-th round, we choose $\mathbf{x}_s(S_{s-1})$ and output $\hat{y}_s = \langle \hat{\mathbf{w}}_{s-1}(S_{s-1}), \mathbf{x}_s(S_{s-1}) \rangle$. Given y_s, we choose k_0 attributes from $\{x_{s,i} : i \notin S_{s-1}\}$. Let $d' = d - k$. We construct $\hat{\mathbf{w}}'_{s-1} \in \mathbb{R}^{d'}$ by removing the elements $\hat{w}_{s,i}$ from $\hat{\mathbf{w}}_{s-1}$ for all $i \in S_{s-1}$. Then we define $\bar{\mathbf{q}}_s = \frac{1}{\|\hat{\mathbf{w}}'_{s-1}\|_1}(|\hat{w}'_{s-1,1}|, \ldots, |\hat{w}'_{s-1,d'}|)$, and send $(k_0, d', \hat{\mathbf{w}}'_{s-1})$ into Algorithm 2 that returns a set $B'_s \subseteq [d]$. Let $B_s = B'_s \cup S_{s-1}$. We can construct $\mathbf{H}_{\mathcal{I}_s}$ and $\hat{\mathbf{X}}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s}$ following DS-OSLRC, in which $\mathcal{I}_s = \{1, 2, \ldots, s\}$. Let $g_{d',k_0} = \frac{(d'-1)(d'-2)}{(k_0-1)(k_0-2)}$, and a_1, a_2, a_3 follow the definition in (9). Let

$$\begin{cases} \mu_1 = \dfrac{9}{9-2\sqrt{3}}, \quad \mu_2 = \dfrac{1}{1 - \frac{\sqrt{6}}{9\sqrt{\frac{d'-2}{k_0-2}\ln\frac{d^2}{\delta}}}}, \\ s_0 = \dfrac{24^2 \cdot k^2 g_{d',k_0}}{\delta_S^4} \ln\dfrac{d^2}{\delta}, \quad s_1 = \dfrac{24^2 \cdot k^2 g_{d',k_0}}{\delta_S^4} \dfrac{d'-2}{k_0-2} \ln\left(\dfrac{d}{\delta}\right) \ln\dfrac{d^2}{\delta}, \\ a_4 = \delta_S^2 \dfrac{a_1}{k} + 24a_2\sqrt{\dfrac{k_0-2}{d'-2}\ln\dfrac{d^2}{\delta}} + 4a_3\left(24\sqrt{\ln\dfrac{d^2}{\delta}} + \dfrac{\delta_S^2}{k\sqrt{g_{d',k_0}}}\right), \\ a_5 = \dfrac{36}{9-2\sqrt{3}}\left(\delta_S^2\dfrac{2+\sigma}{9k} + \dfrac{8}{\sqrt{3}} + \dfrac{\frac{\sqrt{3}}{9}\delta_S^2}{k\sqrt{g_{d',k_0}\ln\frac{d^2}{\delta}}}\right) + a_2 + \dfrac{2\sqrt{3}a_2}{9-2\sqrt{3}}\sqrt{\dfrac{k_0-2}{(d'-2)\ln\frac{d^2}{\delta}}}. \end{cases} \tag{12}$$

Similar to DS-OSLRC, we define $\hat{\gamma}_s$ as follows,

$$\hat{\gamma}_s = \left(\frac{8}{3} + 2\sigma\right)\frac{g_{d',k_0}}{s}\ln\frac{d}{\delta} + (6.9 + 1.2\sigma)\sqrt{\frac{d'-1}{s(k_0-1)}\ln\frac{d}{\delta}} + \nu_s,$$

in which

$$\nu_s = \begin{cases} \frac{2}{\sqrt{s}}\sqrt{3g_{d',k_0}\ln\frac{d}{\delta}}, & s \leq [1, s_0], \\ \frac{1}{s}\sqrt{3g_{d',k_0}\ln\frac{d}{\delta}}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d',k_0}\ln\frac{d^2}{\delta}} + 2\right), & s = s_0 + 1, \\ \frac{s_0+1}{s}\nu_{s_0+1} + \frac{1}{s}\sqrt{3g_{d',k_0}\ln\frac{d}{\delta}} \cdot \frac{\mu_1 a_4}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s-1}\frac{k^4 g_{d',k_0}^2}{\tau^2}}, & s \in (s_0+1, s_1], \\ \frac{1}{s}\sqrt{3g_{d',k_0}\ln\frac{d}{\delta}}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d',k_0}\ln\frac{d^2}{\delta}} + 2 + \frac{\mu_1 a_4 k^2 g_{d',k_0}}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s_1}\frac{1}{\tau^2}}\right), & s = s_1 + 1, \\ \frac{s_1+1}{s}\nu_{s_1+1} + \frac{1}{s}\sqrt{3g_{d',k_0}\ln\frac{d}{\delta}} \cdot \frac{\mu_2 a_5}{\delta_S^2}\sqrt{\sum_{\tau=s_1+1}^{s-1}\frac{k^2(d'-1)}{\tau(k_0-1)}}, & s > s_1 + 1. \end{cases}$$

Now we can solve $\mathrm{DS}(\hat{\gamma}_s)$ and obtain the solution $\hat{\mathbf{w}}_s$. Different from DS-OSLRC, we do not use ONS to update parameters. There are two reasons. (i) We use $\hat{\mathbf{w}}_{s-1}(S_{s-1})$ to make a prediction at

Algorithm 3: DS-POSLRC

Input: k, k_0, d, δ_S, δ

Initialize $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$, $\hat{\mathbf{x}}_{\mathcal{I}_0} = \mathbf{0}_d$, $\mathbf{H}_{\mathcal{I}_0} = \mathbf{0}_{d\times d}$, S_0;

for $s = 1, 2, \ldots, T$ **do**

  Output the prediction $\hat{y}_s = \langle \hat{\mathbf{w}}_{s-1}(S_{s-1}), \mathbf{x}_s(S_{s-1})\rangle$;

  Obtain $B'_s \subseteq [d]$ from SAMPLING$(k_0, d', \hat{\mathbf{w}}'_{s-1})$;

  Let $B_s = B'_s \cup S_{s-1}$;

  Compute $\hat{\mathbf{X}}_{\mathcal{I}_s}\mathbf{Y}_{\mathcal{I}_s} = \hat{\mathbf{X}}_{\mathcal{I}_{s-1}}\mathbf{Y}_{\mathcal{I}_{s-1}} + \hat{\mathbf{x}}_s y_s$;

  Compute $\mathbf{H}_{\mathcal{I}_s} = \mathbf{H}_{\mathcal{I}_{s-1}} + \mathbf{h}_s$;

  Compute $\hat{\gamma}_s$;

  Output the solution of DS$(\hat{\gamma}_s)$, denoted by $\hat{\mathbf{w}}_s$;

  Select $S_s \subseteq [d]$;

end

each round s, ensuring a tight regret bound. (ii) Although ONS can further improve the regret bound by a factor of $O(\ln T)$, it also introduces additional terms and makes the algorithm more complicate. We name this algorithm DS-POSLRC (Dantzig Selector for POSLR with Compatibility condition) and show the pseudo-code in Algorithm 3.

Lemma 3 (Estimation Error) *Let $S = \mathrm{Supp}(\mathbf{w}^*)$, $\hat{\mathbf{w}}_0 = \frac{1}{d}\mathbf{1}_d$ and $3 \le k_0 = O(k\ln d)$. If $T > (s_1+1)^2$, Assumptions 1-2 hold, and $\mathbf{X}_{\mathcal{I}_s}$ satisfies the $(\delta_S, S, 1)$-compatibility condition for all $s \ge 1$, then with probability at least $1 - T(6+\log_{1.5}\frac{d'+k}{k_0-2})\delta$, DS-POSLRC guarantees, $\forall s \ge 1$,*

$$\|\Delta_s(S)\|_1 \le \begin{cases} \frac{16+8\sigma}{\delta_S^2}\frac{kg_{d',k_0}}{s}\ln\frac{d}{\delta} + \frac{(26+4.8\sigma)k}{\delta_S^2\sqrt{s}}\sqrt{\frac{d'-1}{k_0-1}\ln\frac{d}{\delta}} + \frac{22k}{\delta_S^2}\sqrt{\frac{g_{d',k_0}}{s}\ln\frac{d}{\delta}}, & s \le [1, s_0] \\ \frac{9}{9-2\sqrt{3}}\frac{a_4}{\delta_S^4}\frac{k^2 g_{d',k_0}}{s}, & s \in (s_0, s_1] \\ \mu_2\frac{a_5}{\delta_S^2}\sqrt{\frac{k^2(d'-1)}{s(k_0-1)}}, & s > s_1, \end{cases}$$

in which s_0, s_1, a_4 and a_5 follow the definition in (12).

Proof [of Lemma 3] It can be easily confirmed that Corollary 1 and Corollary 2 are valid with the parameter g_{d',k_0}. The proof of Lemma 3 is same with that of Lemma 2. ■

The algorithm in (Kale et al., 2017) attains $\|\Delta_s\|_1 = O\left(\frac{d}{k_0}\sqrt{\frac{d}{k_0}\frac{k^2}{s}\ln\frac{d}{\delta}}\right)$. For $s \ge s_1$, we improve the convergence rate by a factor of $O(\frac{d}{k_0})$. Next we give the regret bound of DS-POSLRC.

Theorem 2 (Regret Bound w.r.t. $\mathbf{w}^*$) *Let $\varepsilon = k$ and $\delta \in (0,1)$. Under the same assumptions in Lemma 3, with probability at least $1 - (T(6+\log_{1.5}\frac{d+k}{k-2})+1)\delta$, the regret of DS-POSLRC satisfies*

$$\mathrm{Reg}(\mathbf{w}^*) \le \frac{48^2k^2g'_{d',k_0}}{\delta_S^4}\ln\frac{d^2}{\delta} + \frac{2.7a_4^2k^2g_{d',k_0}}{64\delta_S^4\ln\frac{d^2}{\delta}} + 9\mu_2^2a_5^2\frac{k^2(d'-1)}{\delta_S^4(k_0-1)}\ln\frac{T}{s_1+1} + O(1),$$

in which s_1, a_4 and a_5 follow (12) and $O(1)$ hides the lower order constant terms.

Next we compare our regret bound with previous results. Let $\delta = \Theta(\frac{1}{T})$. Then with probability at least $1-\delta$, the regret of DS-POSLRC satsifies

$$\mathrm{Reg}(\mathbf{w}^*) = O\left(\frac{(kd')^2}{\delta_S^4k_0^2}\ln\frac{Td^2}{\delta} + \frac{k^2(d'-1)}{\delta_S^4(k_0-1)}\ln(T)\ln\frac{Td}{\delta}\right).$$

Under RIP, the first algorithm by (Kale et al., 2017) enjoys a regret of $O(\frac{k^2d^3}{k_0^2}\ln(T)\ln\frac{Td}{\delta})$. Our algorithm improves the regret bound by a factor of $O(\min\{\frac{d^2}{k_0^2}, \frac{d}{k_0}\ln T\})$. Besides, our result requires the compatibility condition which is weaker than RIP.

Under the linear independence of features condition, the first algorithm by (Ito et al., 2017) enjoys an expected regret of $O(\frac{d}{\sigma_d^2 k_0}\sqrt{T})$. The regret bound is much worse than ours in terms of the dependence on T. What's more, the linear independence of features condition is stronger than RIP and the compatibility condition.

Remark 2 *It is interesting to explore whether recomputing $\hat{\mathbf{w}}_s$ for $s \in \{2^0, 2^1, 2^2, \ldots\}$ can decrease the per-round time complexity, while maintaining a similar convergence rate and regret bound, as illustrated by the first algorithm in (Kale et al., 2017). In this work, our goal is to demonstrate the power of our algorithm-dependent sampling scheme. It is left as a further work to give more efficient algorithms for (k, k_0, d)-OSLR.*

Appendix B. Technical Lemmas

Lemma 4 *For any two positive integers $2 < a < b$, it must be*

$$\sum_{s=a}^{b}\frac{1}{s^2} \leq \frac{1}{a-1} - \frac{1}{b}, \quad \sum_{s=a}^{b}\frac{1}{bs} \leq \frac{1}{2(a-1)}.$$

Proof [of Lemma 4] As $a > 2$, we have

$$\sum_{s=a}^{b}\frac{1}{s^2} \leq \sum_{s=a}^{b}\frac{1}{s(s-1)} = \sum_{s=a}^{b}\left(\frac{1}{s-1} - \frac{1}{s}\right) = \frac{1}{a-1} - \frac{1}{b}.$$

For the second inequality, we consider two cases.
case 1 $b > 2(a-1)$. If b is even, then by the first inequality, we have

$$\begin{aligned}\sum_{s=a}^{b}\frac{1}{bs} = \sum_{s=a}^{\frac{b}{2}+1}\frac{1}{bs} + \sum_{s=\frac{b}{2}+2}^{b}\frac{1}{bs} \leq \sum_{s=a}^{\frac{b}{2}+1}\frac{1}{2s(s-1)} + \frac{1}{\frac{b}{2}+1} - \frac{1}{b} \leq & \frac{1}{2}\left(\frac{1}{a-1} - \frac{1}{\frac{b}{2}+1}\right) + \frac{1}{2}\frac{1}{\frac{b}{2}+1} \\ = & \frac{1}{2(a-1)}.\end{aligned}$$

If b is odd, then we replace $b/2$ with $(b-1)/2$ in the second term.
case 2 $b \leq 2(a-1)$. By the first inequality in the lemma,

$$\sum_{s=a}^{b}\frac{1}{bs} \leq \sum_{s=a}^{b}\frac{1}{s^2} \leq \frac{1}{a-1} - \frac{1}{b} \leq \frac{1}{2(a-1)},$$

which concludes the proof. ■

Lemma 5 *For any $s \geq 1$, let $\hat{\mathbf{w}}_s$ be the solution of $\mathrm{DS}(\hat{\gamma}_s)$. Let $S_s \subseteq [d]$ satisfy $|S_s| = k$ and for any $i \in S_s$ and $j \in [d] \setminus S_s$, $|\hat{w}_{s,i}| \geq |\hat{w}_{s,j}|$. Let $\mathbf{w}_s = \mathbf{w}^*(S \cap S_s)$. It must be*

$$\|\mathbf{w}_s - \mathbf{w}^*\|_1 \leq \|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1.$$

Proof [of Lemma 5] Unfolding $\|\mathbf{w}_s - \mathbf{w}^*\|_1$ gives

$$\begin{aligned}\|\mathbf{w}_s - \mathbf{w}^*\|_1 = \sum_{i\in S\setminus S_s} |w_i^*| = \sum_{i\in S\setminus S_s} |w_i^* - \hat{w}_{s,i} + \hat{w}_{s,i}| &\le \sum_{i\in S\setminus S_s} |w_i^* - \hat{w}_{s,i}| + \sum_{i\in S\setminus S_s} |\hat{w}_{s,i}| \\ &\le \sum_{i\in S\setminus S_s} |w_i^* - \hat{w}_{s,i}| + \sum_{i\in S_s\setminus S} |\hat{w}_{s,i}| \\ &\le \|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1,\end{aligned}$$

in which $w_i^* = 0$ for all $i \in S_s \setminus S$. We conclude the proof. ■

By Lemma 5, our algorithm can significantly reduce the constant factor on the regret bound. To be specific, the proof of Theorem 1 will make use of the following inequality

$$\|\mathbf{w}_s - \mathbf{w}^*\|_1^2 \le \|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1^2 \overset{\text{Lemma 16}}{\le} 4\|\Delta_s(S)\|_1^2. \tag{13}$$

The second approach to prove Theorem 1 is to use the following inequality (please refer to Lemma 4 in Kale et al. (2017) or Lemma 3 in Ito et al. (2017)),

$$\|\hat{\mathbf{w}}_s(S_s) - \mathbf{w}^*\|_2^2 \le 3\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_2^2. \tag{14}$$

If we define $\mathbf{w}_s = \hat{\mathbf{w}}_s(S_s)$, then by (14) and Lemma 16, we can obtain

$$\|\hat{\mathbf{w}}_s(S_s) - \mathbf{w}^*\|_1^2 \le k\|\hat{\mathbf{w}}_s(S_s) - \mathbf{w}^*\|_2^2 \le 3k\|\Delta_s\|_2^2 \le 3k\|\Delta_s\|_1^2 \le 12k\|\Delta_s(S)\|_1^2.$$

Thus our analysis can reduce the constant factor by a factor of $3k$.

Lemma 6 (Bernstein's inequality for martingale) *Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence w.r.t. the filtration $\mathcal{H} = (\mathcal{H}_k)_{1\le k\le n}$ and with $|X_k| \le a$. Let $Z_t = \sum_{k=1}^t X_k$ be the associated martingale. Denote the sum of the conditional variances by*

$$\Sigma_n^2 = \sum_{k=1}^n \mathbb{E}\left[X_k^2 | \mathcal{H}_{k-1}\right] \le v.$$

Then for all constants $a, v > 0$, with probability at least $1 - \delta$,

$$\max_{t=1,\ldots,n} Z_t < \frac{2}{3}a\ln\frac{1}{\delta} + \sqrt{2v\ln\frac{1}{\delta}}.$$

Lemma 6 is derived from Lemma A.8 in (Cesa-Bianchi and Lugosi, 2006). Note that v must be a constant. Next we give a new Bernstein's inequality for martingale in which v is a random variable depending on $X_1, \ldots, X_n$.

Lemma 7 *Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence w.r.t. the filtration $\mathcal{H} = (\mathcal{H}_k)_{1\le k\le n}$ and with $|X_k| \le a$. Let $Z_t = \sum_{k=1}^t X_k$ be the associated martingale. Denote the sum of the conditional variances by $\Sigma_n^2 = \sum_{k=1}^n \mathbb{E}\left[X_k^2|\mathcal{H}_{k-1}\right] \le v$, where $v \in [a_1 n, a_2 n]$ is a random variable depending on $X_1, \ldots, X_n$ and $0 < a_1 < a_2$ are constants. Then for any constant $a > 0$, with probability at least $1 - \left(1 + \log_\beta \frac{a_2}{a_1}\right)\delta$,*

$$\max_{t=1,\ldots,n} Z_t < \frac{2}{3}a\ln\frac{1}{\delta} + \sqrt{2\beta v\ln\frac{1}{\delta}},$$

in which $\beta > 1$ is a constant.

Lemma 7 is a slight variant of Lemma 1 in (Li et al., 2024).

Proof [of Lemma 7] We divide the interval $[a_1 n, a_2 n]$ as follows

$$[a_1 n, a_2 n] \subseteq \bigcup_{j=0}^{\lfloor \log_\beta \frac{a_2}{a_1} \rfloor} \left[a_1 n \cdot \beta^j, a_1 n \cdot \beta^{j+1}\right).$$

We decompose the random event as follows,

$$\begin{aligned}
&\mathbb{P}\left[\max_{t=1,\ldots,n} Z_t > \frac{2}{3} a \ln\frac{1}{\delta} + \sqrt{2\beta v \ln\frac{1}{\delta}}, \Sigma_n^2 \le v\right] \\
=&\mathbb{P}\left[\max_{t\le n} Z_t > \frac{2}{3} a \ln\frac{1}{\delta} + \sqrt{2\beta v \ln\frac{1}{\delta}}, \Sigma_n^2 \le v, \cup_{j=0}^{\lfloor \log_\beta \frac{a_2}{a_1} \rfloor} a_1 n \cdot \beta^j \le v < a_1 n \cdot \beta^{j+1}\right] \\
\le& \sum_{j=0}^{\lfloor \log_\beta \frac{a_2}{a_1} \rfloor} \mathbb{P}\left[\max_{t\le n} Z_t > \frac{2}{3} a \ln\frac{1}{\delta} + \sqrt{2\beta v \ln\frac{1}{\delta}}, \Sigma_n^2 \le v, a_1 n \cdot \beta^j \le v < a_1 n \cdot \beta^{j+1}\right] \\
\le& \sum_{j=0}^{\lfloor \log_\beta \frac{a_2}{a_1} \rfloor} \mathbb{P}\left[\max_{t\le n} Z_t > \frac{2}{3} a \ln\frac{1}{\delta} + \sqrt{2\beta \cdot a_1 n \cdot \beta^j \ln\frac{1}{\delta}}, \Sigma_n^2 \le v, a_1 n \cdot \beta^j \le v < a_1 n \cdot \beta^{j+1}\right] \\
=& \sum_{j=0}^{\lfloor \log_\beta \frac{a_2}{a_1} \rfloor} \mathbb{P}\left[\max_{t\le n} Z_t > \frac{2}{3} a \ln\frac{1}{\delta} + \sqrt{2 \cdot a_1 n \cdot \beta^{j+1} \ln\frac{1}{\delta}}, \Sigma_n^2 \le v, a_1 n \cdot \beta^j \le v < a_1 n \cdot \beta^{j+1}\right] \\
\le& \left(1 + \log_\beta \frac{a_2}{a_1}\right)\delta,
\end{aligned}$$

in which we use Lemma 6 for each sub-event. ■

Lemma 8 (Hazan et al. (2007)) *Let $\mathbf{u}_t \in \mathbb{R}^n$, for $t = 1, 2, \ldots, T$, be a sequence of vectors such that for some $r > 0$, $\|\mathbf{u}_t\|_2 \le r$. Define $\mathbf{V}_t = \sum_{s=1}^t \mathbf{u}_s \mathbf{u}_s^\top + \varepsilon \cdot \mathbf{I}_{n\times n}$. Then*

$$\sum_{t=1}^T \mathbf{u}_t^\top \mathbf{V}_t^{-1} \mathbf{u}_t \le n \ln\left(\frac{r^2 T}{\varepsilon} + 1\right).$$

Lemma 9 *For each $s = 1, 2, \ldots$, let $\mathbf{q}_s = (\frac{|\hat{w}_{s-1,1}|}{\|\hat{\mathbf{w}}_{s-1}\|_1}, \ldots, \frac{|\hat{w}_{s-1,d}|}{\|\hat{\mathbf{w}}_{s-1}\|_1})$ and B_s be the indexes of the features selected by DS-OSLRC.*

$$\begin{aligned}
\forall i \in [d], \quad \mathbb{P}[i \in B_s] =& \frac{d-k}{d-1} q_{s,i} + \frac{k-1}{d-1}, \\
\forall i \ne j \in [d], \quad \mathbb{P}[i, j \in B_s] =& \frac{1}{g_{d,k}} + \frac{1}{g_{d,k}} \cdot (q_{s,i} + q_{s,j}), \\
\forall i \ne j \ne r \in [d], \quad \mathbb{P}[i, j, r \in B_s] =& \frac{1}{g_{d,k}} \cdot \frac{k-3}{d-3} + \frac{1}{g_{d,k}} \cdot \frac{d-k}{d-3} \cdot (q_{t,i} + q_{t,j} + q_{t,r}).
\end{aligned} \tag{15}$$

Besides,

$$\mathbb{E}_s\left[\hat{\mathbf{x}}_{s^2}\right] = \mathbf{x}_{s^2}, \quad \mathbb{E}_s\left[\mathbf{h}_{s^2}\right] = \mathbf{x}_{s^2} \mathbf{x}_{s^2}^\top,$$

where $\mathbb{E}_s[\cdot] = \mathbb{E}[\cdot | B_1, \ldots, B_{s-1}]$ is the conditional expectation and is taken with respect to B_s.

(m,n)	1	2	3	...	k
1	-	$p_{1,2}[i,j]$	$p_{1,3}[i,j]$	...	$p_{1,k}[i,j]$
2	$p_{2,1}[i,j]$	-	$p_{2,3}[i,j]$	...	$p_{2,k}[i,j]$
3	$p_{3,1}[i,j]$	$p_{3,2}[i,j]$	-	...	$p_{3,k}[i,j]$
...	...	...	...	...	...
k	$p_{k,1}[i,j]$	$p_{k,2}[i,j]$	$p_{k,3}[i,j]$	...	-

Table 4: The probabilities of the event $(i,j) \in B_s$. The notation "-" means the event is invalid.

Proof [of Lemma 9] It is easy to prove the first equality in (15). For any $i \in B_s$, summing all the probabilities selected at each sampling step $n = 1, 2, 3, \ldots, k$, yields the following result.

$$\begin{aligned}\mathbb{P}[i \in B_s] =& p_{s,i} + (1-p_{s,i}) \cdot \frac{1}{d-1} + \sum_{n=3}^{k} \underbrace{(1-p_{s,i}) \cdot \prod_{r=1}^{n-2} \left(1 - \frac{1}{d-r}\right) \frac{1}{d-(n-1)}}_{\mathbb{P}[i \text{ is selected at the } n\text{-th round}]} \\ =& p_{s,i} + (1-p_{s,i}) \cdot \frac{k-1}{d-1}.\end{aligned}$$

Rearranging terms recoveries the first equality. It is more complicated to prove the second equality and the third equality in (15). We consider any pair of (i,j), and analyze the probability $\mathbb{P}[i \neq j \in B_s]$. For clarity, we enumerate all of the combinations (m,n) where the i-th feature is selected during the m-th sampling and the j-th feature is selected during the n-th sampling in Table 4, in which $p_{m,n}[i,j]$ is the corresponding probability. It is obvious that $\mathbb{P}[i \neq j \in B_s] = \sum_{m \neq n} p_{m,n}[i,j]$. Next we analyze $p_{n,m}[i,j]$.

It is worth mentioning that the probability that the i-th feature is selected at the first sampling does not equal to the probability that the j-th feature is selected at the first sampling. Thus we must separately analyze the cases $m = 1$ and $n = 1$. For $m \geq 2$ and $n \geq 2$, the probability that the i-th feature selected at the r-th sampling is same with that of the j-th feature, in which $r \geq 2$. Thus we just analyze $p_{m,n}$ for all $m = 2, 3, \ldots$, which equals to $p_{m,n}$ for all $n = 2, 3, \ldots$ by the symmetry.

We first consider the case that the i-th feature is selected at the first sampling, i.e., $m = 1$.

$$\begin{aligned} p_{1,2}[i,j] =& q_{s,i} \frac{1}{d-1}, \\ \forall n \geq 3, \quad p_{1,n}[i,j] =& q_{s,i} \cdot \prod_{r=1}^{n-2} \left(1 - \frac{1}{d-r}\right) \frac{1}{d-(n-1)} = q_{s,i} \frac{1}{d-1}. \end{aligned}$$

If the j-th feature is sampled at the first sampling, i.e., $n = 1$, then we have

$$\begin{aligned} p_{2,1}[i,j] =& q_{s,j} \frac{1}{d-1}, \\ \forall m \geq 3, \quad p_{m,1}[i,j] =& q_{s,j} \cdot \prod_{r=1}^{m-2} \left(1 - \frac{1}{d-r}\right) \frac{1}{d-(m-1)} = q_{s,j} \frac{1}{d-1}. \end{aligned}$$

Then we consider the case $m = 2$.

$$\begin{aligned} p_{2,3}[i,j] =&(1-q_{s,i}-q_{s,j})\cdot\frac{1}{d-1}\cdot\frac{1}{d-2},\\ \forall n\geq 4,\quad p_{2,n}[i,j] =&(1-q_{s,i}-q_{s,j})\cdot\frac{1}{d-1}\cdot\prod_{r=2}^{n-2}\left(1-\frac{1}{d-r}\right)\cdot\frac{1}{d-(n-1)}\\ =&\frac{1-q_{s,i}-q_{s,j}}{(d-1)(d-2)}. \end{aligned}$$

By the symmetry, the probability of sampling i and j for $m \geq 2$ and $n = 2$ is

$$\forall m\geq 3\quad p_{m,2}[i,j] = (1-q_{s,i}-q_{s,j})\cdot\frac{1}{d-1}\cdot\frac{1}{d-2}.$$

Finally, we consider the case $m \geq 3$.

$$\begin{aligned} p_{m,m+1}[i,j] =&(1-q_{s,i}-q_{s,j})\cdot\prod_{r=1}^{m-2}\left(1-\frac{2}{d-r}\right)\frac{1}{d-(m-1)}\cdot\frac{1}{d-m}\\ =&\frac{1-q_{s,i}-q_{s,j}}{(d-1)(d-2)},\\ \forall n\geq m+2,\quad p_{m,n}[i,j] =&(1-q_{s,i}-q_{s,j})\prod_{r=1}^{m-2}\left(1-\frac{2}{d-r}\right)\frac{1}{d-(m-1)}\prod_{r=0}^{n-m-2}\frac{1-\frac{1}{d-m-r}}{d-(n-1)}\\ =&\frac{1-q_{s,i}-q_{s,j}}{(d-1)(d-2)}. \end{aligned}$$

By the symmetry, for the case of $n \geq 3$, we have

$$\forall m\geq n+1\quad p_{m,n}[i,j] = (1-q_{s,i}-q_{s,j})\cdot\frac{1}{d-1}\cdot\frac{1}{d-2}.$$

Combining all of the above results yields

$$\begin{aligned} \mathbb{P}[i\neq j\in B_s] =&\sum_{n\neq 1}p_{1,n}[i,j]+\sum_{m=2}^{k}\sum_{n\neq m}p_{m,n}[i,j]\\ =&q_{s,i}\cdot\frac{k-1}{d-1}+\sum_{m=2}^{k}\left(\frac{q_{s,j}}{d-1}+(1-q_{s,i}-q_{s,j})\frac{k-2}{(d-1)(d-2)}\right)\\ =&q_{s,i}\cdot\frac{k-1}{d-1}+(k-1)\cdot\left(\frac{q_{s,j}}{d-1}+(1-q_{s,i}-q_{s,j})\frac{k-2}{(d-1)(d-2)}\right)\\ =&\frac{(k-1)(k-2)}{(d-1)(d-2)}+\frac{(k-1)(d-k)}{(d-1)(d-2)}\cdot(q_{s,i}+q_{s,j}), \end{aligned}$$

which recoveries the second equality in (15).

Finally, we will prove the third equality. We consider any pair of (i, j, r), and analyze $\mathbb{P}[i \neq j \neq r \in B_t]$. We can enumerate all combinations (m, n, o) where the i-th feature is selected during the m-th sampling, the j-th feature is selected during the n-th sampling, and the r-th feature is selected during the o-th sampling. We consider four cases.

- $m = 1$
  Assuming that the i-th feature was selected at the first sampling step. We only need to compute $\mathbb{P}[j \neq r \in B_s]$ following the second equality in (15). To be specific, we will sample $k-1$ indexes from $[d] \setminus \{m\}$ without replacement. Thus we have

$$\mathbb{P}[i \neq j \neq r \in B_s] = q_{s,i} \cdot \left(\frac{(k-2)(k-3)}{(d-2)(d-3)} + \frac{(k-2)(d-k)}{(d-2)(d-3)} \left(\frac{1}{d-1} + \frac{1}{d-1} \right) \right),$$

  in which we use $(d-1, k-1, \frac{1}{d-1}, \frac{1}{d-1})$ to replace the value of $(d, k, q_{t,i}, q_{t,j})$.

- $n = 1$
  The analysis is same with that of $m = 1$.

$$\mathbb{P}[i \neq j \neq r \in B_s] = q_{s,j} \cdot \left(\frac{(k-2)(k-3)}{(d-2)(d-3)} + \frac{(k-2)(d-k)}{(d-2)(d-3)} \cdot \frac{2}{d-1} \right).$$

- $o = 1$
  The analysis is also same with that of $m = 1$.

$$\mathbb{P}[i \neq j \neq r \in B_s] = q_{s,r} \cdot \left(\frac{(k-2)(k-3)}{(d-2)(d-3)} + \frac{(k-2)(d-k)}{(d-2)(d-3)} \cdot \frac{2}{d-1} \right).$$

- $m \neq 1, n \neq 1, o \neq 1$:
  Assuming that the u-th feature satisfying $u \neq i, u \neq j, u \neq r$ has been selected. Then we will sample $k-1$ indexes from $[d] \setminus \{m\}$ without replacement. By the analyzing of the second equality in (15), it is easy to be verified that the probabilities that any combination of (m, n, o) for $m \geq 2, n \geq 2, o \geq 2$ are the same. In this case, the number of combinations of (m, n, o) is $(k-1)(k-2)(k-3)$. Without loss of generality, assuming that $2 = m < n < o$. Thus

$$\begin{aligned}
&\mathbb{P}[i \neq j \neq r \in B_s] \\
&= (1 - q_{s,i} - q_{s,j} - q_{s,r})\,(k-1)(k-2)(k-3) \left(\frac{(k-2)(d-k)}{(d-2)(d-3)} + \frac{(k-2)(d-k)}{(d-2)(d-3)} \frac{2}{d-1} \right) \\
&= (1 - q_{s,i} - q_{s,j} - q_{s,r}) \cdot (k-1)(k-2)(k-3) \cdot \left(\frac{1}{d-1} \cdot \frac{1}{d-2} \cdot \frac{1}{d-3} \right).
\end{aligned}$$

Summing all results gives

$$\begin{aligned}
&\mathbb{P}[i \neq j \neq r \in B_s] \\
=&\left(\frac{(k-2)(k-3)}{(d-2)(d-3)} + \frac{(k-2)(d-k)}{(d-2)(d-3)} \cdot \frac{2}{d-1} \right) (q_{s,i} + q_{s,j} + q_{s,r}) + \\
&(1 - q_{s,i} - q_{s,j} - q_{s,r}) \cdot \frac{(k-1)(k-2)(k-3)}{(d-1)(d-2)(d-3)} \\
=&\frac{(k-1)(k-2)}{(d-1)(d-2)} \cdot (q_{s,i} + q_{s,j} + q_{s,r}) + (1 - q_{s,i} - q_{s,j} - q_{s,r}) \cdot \frac{(k-1)(k-2)(k-3)}{(d-1)(d-2)(d-3)} \\
=&\frac{(k-1)(k-2)(k-3)}{(d-1)(d-2)(d-3)} + \frac{(k-1)(k-2)(d-k)}{(d-1)(d-2)(d-3)} \cdot (q_{s,i} + q_{s,j} + q_{s,r}),
\end{aligned}$$

which concludes the proof.

As B_s depends on $B_1, B_2, \ldots, B_{s-1}$, thus $\hat{\mathbf{w}}_0$, $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_s$ are not independent. Let $\mathbb{E}_s[\cdot] = \mathbb{E}[\cdot|B_1, \ldots, B_{s-1}]$ be the conditional expectation. It is easy to show that $\mathbb{E}_s[\hat{x}_{s^2,i}] = x_{s^2,i}$ for all $i \in [d]$, implying $\mathbb{E}_s[\hat{\mathbf{x}}_{s^2}] = \mathbf{x}_{s^2}$. Similarly, it must be $\mathbb{E}_s[h_{s^2}[i,i]] = x_{s^2,i}^2$ and $\mathbb{E}_s[h_{s^2}[i,j]] = x_{s^2,i}x_{s^2,j}$ for all $i \neq j$. Thus $\mathbb{E}_s[\mathbf{h}_{s^2}] = \mathbf{x}_{s^2}\mathbf{x}_{s^2}^\top$. ■

Lemma 10 *For any $a > 0, b > 0, c > 0, d > 0$ and $0 \leq x \leq y \leq 1$, if $a + b = c + d = 1$ and $a \leq c$, then it must be*

$$\frac{a + bx}{c + dy} \leq \frac{a + b}{c + d}.$$

Proof [of Lemma 10] The above inequality is equivalent to

$$\begin{aligned}
&(a + bx)(c + d) \leq (a + b)(c + dy)\\
\Leftrightarrow &ad + bcx + bdx \leq ady + bc + bdy\\
\Leftrightarrow &a(1 - c) + (1 - a)cx + (1 - a)(1 - c)x \leq a(1 - c)y + (1 - a)c + (1 - a)(1 - c)y\\
\Leftrightarrow &a(1 - c)(1 - y) \leq c(1 - a)(1 - x) + (1 - a)(1 - c)(y - x).
\end{aligned}$$

By applying the constraints on these variables, the following two inequalities can be derived.

$$1 - y \leq 1 - x, \quad 1 - c \leq 1 - a.$$

It is obvious that the inequality in Lemma 10 holds. ■

Lemma 11 *Let $k \geq 3$. At any $\tau \geq 1$, for any $i \neq j \neq r \in [d]$, it must be*

$$\frac{\mathbb{P}[i, j, r \in B_\tau]}{\mathbb{P}[i, j \in B_\tau] \cdot \mathbb{P}[i, r \in B_\tau]} \leq \frac{d - 1}{k - 1}.$$

Proof [of Lemma 11] By Lemma 9 and Lemma 10, we can obtain

$$\begin{aligned}
&\frac{\mathbb{P}[i, j, r \in B_\tau]}{\mathbb{P}[i, j \in B_\tau] \cdot \mathbb{P}[i, r \in B_\tau]}\\
=&\frac{\frac{(k-1)(k-2)(k-3)}{(d-1)(d-2)(d-3)} + \frac{(k-1)(k-2)(d-k)}{(d-1)(d-2)(d-3)} \cdot (q_{\tau-1,i} + q_{\tau-1,j} + q_{\tau-1,r})}{\left(\frac{(k-1)(k-2)}{(d-1)(d-2)} + \frac{(k-1)(d-k)}{(d-1)(d-2)} \cdot (q_{\tau-1,i} + q_{\tau-1,j})\right) \cdot \left(\frac{(k-1)(k-2)}{(d-1)(d-2)} + \frac{(k-1)(d-k)}{(d-1)(d-2)} \cdot (q_{\tau-1,i} + q_{\tau-1,r})\right)}\\
=&\frac{d - 1}{k - 1} \cdot \frac{\frac{k-3}{d-3} + \frac{d-k}{d-3} \cdot (q_{\tau-1,i} + q_{\tau-1,j} + q_{\tau-1,r})}{\frac{k-2}{d-2} + \frac{d-k}{d-2}(2q_{\tau-1,i} + q_{\tau-1,j} + q_{\tau-1,r}) + \frac{(d-k)^2}{(d-2)(k-2)}(q_{\tau-1,i} + q_{\tau-1,j})(q_{\tau-1,i} + q_{\tau-1,r})}\\
\leq&\frac{d - 1}{k - 1} \cdot \frac{\frac{k-3}{d-3} + \frac{d-k}{d-3}}{\frac{k-2}{d-2} + \frac{d-k}{d-2}} = \frac{d - 1}{k - 1},
\end{aligned}$$

which concludes the proof. ■

Lemma 12 *For any $\tau \geq 1$ and $\mathbf{v} \in \mathbb{R}^d$, let $\mathbf{z}_\tau = \mathbf{h}_{\tau^2}\mathbf{v} - \mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top \mathbf{v}$. Then*

$$\begin{aligned}\forall i \in [d], \quad \mathbb{E}_\tau[z_{\tau,i}] =& 0,\\ \mathbb{E}_\tau\left[z_{\tau,i}^2\right] =& \left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^4 \cdot v_i^2 + 2\sum_{j\neq i}\left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^3 x_{\tau^2,j} \cdot v_i v_j + \\ & \sum_{j\neq i}\sum_{r\neq i}\left(\frac{\mathbb{P}[i,j,r \in B_\tau]}{\mathbb{P}[i,j \in B_\tau]\cdot\mathbb{P}[i,r \in B_\tau]} - 1\right) x_{\tau^2,i}^2 x_{\tau^2,j} x_{\tau^2,r} \cdot v_r v_j.\end{aligned}$$

Proof [of Lemma 12] Substituting into the definition of $\mathbf{h}_{\tau^2}$, and using Lemma 9, it is easy to verify that $\mathbb{E}_\tau[z_{\tau,i}] = 0$ for all $i \in [d]$. It is more challenge to analyze the variance, as the even $i \in B_\tau$, $i \neq j \in B_\tau$ and $i \neq j \neq r \in B_\tau$ are not independent. Note that we have $\mathbb{I}_{i\in B_\tau} \cdot \mathbb{I}_{i,j\in B_\tau} = \mathbb{I}_{i,j\in B_\tau}$ and $\mathbb{I}_{i,j\in B_\tau} \cdot \mathbb{I}_{i,r\in B_\tau} = \mathbb{I}_{i,j,r\in B_\tau}$.

$$\begin{aligned}&\mathbb{E}_\tau[z_{\tau,i}^2]\\ =&\mathbb{E}_\tau\left[\left(\frac{x_{\tau^2,i}^2}{\mathbb{P}[i \in B_\tau]}\mathbb{I}_{i\in B_\tau} - x_{\tau^2,i}^2\right)^2 v_i^2 + \left(\sum_{j\neq i}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j \in B_\tau]}\mathbb{I}_{i,j\in B_\tau} - x_{\tau^2,i}x_{\tau^2,j}\right) v_j\right)^2\right] +\\ &2\mathbb{E}_\tau\left[\left(\frac{x_{\tau^2,i}^2}{\mathbb{P}[i \in B_\tau]}\mathbb{I}_{i\in B_\tau} - x_{\tau^2,i}^2\right)\cdot v_i\left(\sum_{j\neq i}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j \in B_\tau]}\mathbb{I}_{i,j\in B_\tau} - x_{\tau^2,i}x_{\tau^2,j}\right) v_j\right)\right]\\ =&\left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^4 \cdot v_i^2 + 2\sum_{j\neq i}\left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^3 x_{\tau^2,j} \cdot v_i v_j +\\ &\mathbb{E}_\tau\left[\sum_{j\neq i}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j \in B_\tau]}\mathbb{I}_{i,j\in B_\tau} - x_{\tau^2,i}x_{\tau^2,j}\right)^2 v_j^2\right] +\\ &\mathbb{E}_\tau\left[\sum_{j\neq i}\sum_{r\neq i,r\neq j}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j \in B_\tau]}\mathbb{I}_{i,j\in B_\tau} - x_{\tau^2,i}x_{\tau^2,j}\right)\left(\frac{x_{\tau^2,i}x_{\tau^2,r}}{\mathbb{P}[i,r \in B_\tau]}\mathbb{I}_{i,r\in B_\tau} - x_{\tau^2,i}x_{\tau^2,r}\right) v_r v_j\right]\\ =&\left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^4 \cdot v_i^2 + 2\sum_{j\neq i}\left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^3 x_{\tau^2,j} \cdot v_i v_j +\\ &\sum_{j\neq i}\sum_{r\neq i}\left(\frac{\mathbb{P}[i,j,r \in B_\tau]}{\mathbb{P}[i,j \in B_\tau]\cdot\mathbb{P}[i,r \in B_\tau]} - 1\right) x_{\tau^2,i}^2 x_{\tau^2,j} x_{\tau^2,r} \cdot v_r v_j.\end{aligned}$$

We conclude the proof. ■

By Lemma 12, we give two types of upper bounds on the variance.

Corollary 1 *Assuming that $\|\mathbf{x}_t\|_\infty \leq 1, t \in [T]$. For any $\tau \geq 1$, let $\mathbf{v} = \hat{\mathbf{w}}_{\tau-1}$ in Lemma 12. Then*

$$\begin{aligned}\forall i \in [d], \quad |z_{\tau,i}| \leq& \frac{(d-1)(d-2)}{(k-1)(k-2)}\|\hat{\mathbf{w}}_{\tau-1}\|_1,\\ \mathbb{E}_\tau\left[z_{\tau,i}^2\right] \leq& \frac{2(d-1)}{k-1}\|\hat{\mathbf{w}}_{\tau-1}\|_1^2.\end{aligned}$$

Proof [of Corollary 1] By Lemma 9, we can obtain

$$\begin{aligned}|z_{\tau,i}| =& \left|\sum_{j=1}^{d}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j\in B_\tau]}\mathbb{I}_{i,j\in B_\tau} - x_{\tau^2,i}x_{\tau^2,j}\right)\hat{w}_{\tau-1,j}\right| \\ \leq& \left|\left(\frac{d-1}{k-1}-1\right)|\hat{w}_{\tau-1,i}| + \sum_{j\neq i}\left(\frac{(d-1)(d-2)}{(k-1)(k-2)}-1\right)\cdot|\hat{w}_{\tau-1,j}|\right| \\ \leq& \frac{(d-1)(d-2)}{(k-1)(k-2)}\|\hat{\mathbf{w}}_{\tau-1}\|_1.\end{aligned}$$

Next we analyze the second-order moment. It is easy to show that the second-order moment is $O(\frac{d-1}{k-1})$. The technical challenge lies in maintaining a small constant. To this end, we will use Lemma 11. By Lemma 12, we just need to analyze the following three terms.

$$\begin{aligned}\left(\frac{1}{\mathbb{P}[i\in B_\tau]}-1\right)x_{\tau^2,i}^4\cdot v_i^2 \leq& \frac{1}{\frac{k-1}{(d-1)q_{\tau,i}}+\frac{d-k}{d-1}}|\hat{w}_{\tau-1,i}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 - |\hat{w}_{\tau-1,i}|^2, \\ \leq& |\hat{w}_{\tau-1,i}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 - |\hat{w}_{\tau-1,i}|^2. \\ \left(\frac{1}{\mathbb{P}[i\in B_\tau]}-1\right)x_{\tau^2,i}^3x_{\tau^2,j}\cdot v_iv_j \leq& |\hat{w}_{\tau-1,j}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 - |\hat{w}_{\tau-1,i}|\cdot|\hat{w}_{\tau-1,j}|.\end{aligned}$$

If $r=j$, then we have

$$\begin{aligned}\left(\frac{\mathbb{P}[i,j,r\in B_\tau]}{\mathbb{P}[i,j\in B_\tau]\cdot\mathbb{P}[i,r\in B_\tau]}-1\right)x_{\tau^2,i}^2x_{\tau^2,j}x_{\tau^2,r}v_rv_j =& \left(\frac{1}{\mathbb{P}[i,j\in B_\tau]}-1\right)x_{\tau^2,i}^2x_{\tau^2,j}x_{\tau^2,r}v_j^2 \\ \leq& \frac{1}{\frac{(k-1)(k-2)}{(d-1)(d-2)q_{\tau,j}}+\frac{(k-1)(d-k)}{(d-1)(d-2)}}\cdot|\hat{w}_{\tau-1,j}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 \\ \leq& \frac{d-1}{k-1}\cdot|\hat{w}_{\tau-1,j}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1.\end{aligned}$$

If $r\neq j$, then by Lemma 11, we have

$$\left(\frac{\mathbb{P}[i,j,r\in B_\tau]}{\mathbb{P}[i,j\in B_\tau]\cdot\mathbb{P}[i,r\in B_\tau]}-1\right)x_{\tau^2,i}^2x_{\tau^2,j}x_{\tau^2,r}\cdot v_rv_j \leq \left(\frac{d-1}{k-1}-1\right)\cdot|\hat{w}_{\tau-1,r}|\cdot|\hat{w}_{\tau-1,j}|.$$

Summing the above results yields

$$\begin{aligned}\mathbb{E}_\tau\left[z_{\tau,i}^2\right] \leq& |\hat{w}_{\tau-1,i}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 + 2\sum_{j\neq i}|\hat{w}_{\tau-1,j}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 - \|\hat{\mathbf{w}}_{\tau-1}\|_1^2 + \\ & \frac{d-1}{k-1}\sum_{j\neq i}|\hat{w}_{\tau-1,j}|\cdot\|\hat{\mathbf{w}}_{\tau-1}\|_1 + \left(\frac{d-1}{k-1}-1\right)\sum_{j\neq i}\sum_{r\neq i,r\neq j}|\hat{w}_{\tau-1,r}|\cdot|\hat{w}_{\tau-1,j}| \\ =& \frac{2(d-1)}{k-1}\|\hat{\mathbf{w}}_{\tau-1}\|_1^2,\end{aligned}$$

which concludes the proof. ■

Corollary 2 *Assuming that* $\|\mathbf{x}_t\|_\infty \leq 1, t \in [T]$. *Let* $\mathbf{v}$ *be any vector in* $\mathbb{R}^d$ *in Lemma 12. Then*

$$\forall i \in [d], \quad |z_{\tau,i}| \leq \frac{(d-1)(d-2)}{(k-1)(k-2)} \|\mathbf{v}\|_1,$$
$$\mathbb{E}_\tau \left[z_{\tau,i}^2\right] \leq \left(\frac{(d-1)(d-2)}{(k-1)(k-2)} - 1\right) \|\mathbf{v}\|_2^2 + \left(\frac{d-1}{k-1} - 1\right) \left(\|\mathbf{v}\|_1^2 - \|\mathbf{v}\|_2^2\right).$$

Proof [of Corollary 2] The analysis on $|z_{\tau,i}|$ is same with that of Corollary 1. By Lemma 11,

$$\begin{aligned}
\mathbb{E}_\tau \left[z_{\tau,i}^2\right] = & \left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^4 \cdot v_i^2 + 2 \sum_{j \neq i} \left(\frac{1}{\mathbb{P}[i \in B_\tau]} - 1\right) x_{\tau^2,i}^3 x_{\tau^2,j} \cdot v_i v_j + \\
& \sum_{j \neq i, r = j} \left(\frac{1}{\mathbb{P}[i,j \in B_\tau]} - 1\right) x_{\tau^2,i}^2 x_{\tau^2,j}^2 \cdot v_j^2 + \\
& \sum_{j \neq i} \sum_{r \neq i, r \neq j} \left(\frac{\mathbb{P}[i,j,r \in B_\tau]}{\mathbb{P}[i,j \in B_\tau] \cdot \mathbb{P}[i,r \in B_\tau]} - 1\right) x_{\tau^2,i}^2 x_{\tau^2,j} x_{\tau^2,r} \cdot v_r v_j \\
\leq & \left(\frac{d-1}{k-1} - 1\right) \left(|v_i|^2 + 2 \sum_{j \neq i} |v_i v_j| + \sum_{j \neq i} \sum_{r \neq i, r \neq j} |v_r v_j|\right) + \sum_{j \neq i, r = j} (g_{d,k} - 1) |v_j|^2 \\
\leq & (g_{d,k} - 1) \|\mathbf{v}\|_2^2 + \left(\frac{d-1}{k-1} - 1\right) \left(\|\mathbf{v}\|_1^2 - \|\mathbf{v}\|_2^2\right),
\end{aligned}$$

which concludes the proof. ■

Next we prove that if γ_s and $\hat{\gamma}_s$ are well selected, then with a high probability, $\mathrm{DS}(\hat{\gamma}_s)$ has a solution at least, denoted by $\hat{\mathbf{w}}_s$ satisfying $\|\hat{\mathbf{w}}_s\|_1 \leq \|\mathbf{w}^*\|_1$.

Lemma 13 *Let* $\delta \in (0,1)$. *For any* $s \geq 1$, *if* $\hat{\gamma}_\tau \geq \gamma_\tau$ *for all* $\tau \leq s$ *where* γ_τ *follows* (6) *and* $\hat{\gamma}_\tau$ *follows* (7), *then*

(i) with probability at least $1 - s\left(5 + \log_{1.5} \frac{2(d-2)}{3(k-2)}\right)\delta$, $\frac{1}{\tau} \left\|\hat{\mathbf{X}}_{\mathcal{I}_\tau} \mathbf{Y}_{\mathcal{I}_\tau} - \mathbf{H}_{\mathcal{I}_\tau} \mathbf{w}^*\right\|_\infty \leq \gamma_\tau$ *for all* $\tau \leq s$,

(ii) $\mathrm{DS}(\hat{\gamma}_\tau)$ *is feasible and its optimal solution* $\hat{\mathbf{w}}_\tau$ *satisfying* $\|\hat{\mathbf{w}}_\tau\|_1 \leq \|\mathbf{w}^*\|_1$ *for all* $\tau \leq s$.

Proof [of Lemma 13] The main challenge to prove this lemma is the coupling between (i) and (ii). To address this issue, we will use an induction method to prove this lemma.

First, we consider $s = 1$. In this case,

$$\gamma_1 = \left(\frac{8}{3} + 2\sigma\right) g_{d,k} \ln \frac{d}{\delta} + (6.9 + 1.2\sigma) \sqrt{\frac{d-1}{k-1} \ln \frac{d}{\delta}}, \quad \hat{\mathbf{w}}_0 = \frac{1}{d} \mathbf{1}_d.$$

Recalling that $y_1 = \langle \mathbf{w}^*, \mathbf{x}_1 \rangle + \eta_1$. With probability at least $1-\delta$,

$$\begin{aligned}
\|\mathbf{h}_1\mathbf{w}^* - \hat{\mathbf{x}}_1 y_1\|_\infty &= \left\|\mathbf{h}_1\mathbf{w}^* - \hat{\mathbf{x}}_1\left(\mathbf{x}_1^\top\mathbf{w}^* + \eta_1\right)\right\|_\infty \\
&= \left\|\left(\mathbf{h}_1 - \mathbf{x}_1\mathbf{x}_1^\top\right)\mathbf{w}^* + \left(\mathbf{x}_1\mathbf{x}_1^\top - \hat{\mathbf{x}}_1\mathbf{x}_1^\top\right)\mathbf{w}^* - \hat{\mathbf{x}}_1\eta_1\right\|_\infty \\
&\le \left\|\left(\mathbf{h}_1 - \mathbf{x}_1\mathbf{x}_1^\top\right)\mathbf{w}^*\right\|_\infty + \left\|\left(\mathbf{x}_1 - \hat{\mathbf{x}}_1\right)\mathbf{x}_1^\top\mathbf{w}^*\right\|_\infty + \|\hat{\mathbf{x}}_1\eta_1\|_\infty \\
&\le \max_{i\in[d]}\left|\sum_{j=1}^{d}(h_1[i,j] - x_{1,i}x_{1,j})w_j^*\right| + \frac{d-1}{k-1} + \|\hat{\mathbf{x}}_1\|_\infty\cdot\eta_1 \\
&\le \frac{(d-1)(d-2)}{(k-1)(k-2)} + \frac{d-1}{k-1} + \frac{d-1}{k-1}\cdot\sigma\sqrt{2\ln\frac{1}{\delta}} < \gamma_1.
\end{aligned}$$

For the Gaussian random variable $\eta_1 \sim \mathcal{N}(0,\sigma^2)$, with probability (w.p.) at least $1-\delta$, it must be $|\eta_1| \le \sigma\sqrt{2\ln\frac{1}{\delta}}$. We will give a detail proof later. Since $\gamma_1 \le \hat{\gamma}_1$, $\mathrm{DS}(\hat{\gamma}_1)$ has a solution at least, i.e., $\hat{\mathbf{w}}_1 = \mathbf{w}^*$. This lemma holds for $s=1$.

Assuming (i) holds for any $s = r-1 \ge 1$, that is, w.p. at least $1-(r-1)\left(5+\log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$, $\frac{1}{\tau}\left\|\hat{\mathbf{X}}_{\mathcal{I}_\tau}\mathbf{Y}_{\mathcal{I}_\tau} - \mathbf{H}_{\mathcal{I}_\tau}\mathbf{w}^*\right\|_\infty \le \gamma_\tau$ for all $\tau \le r-1$. Since $\hat{\gamma}_\tau \ge \gamma_\tau$ for all $\tau \le r-1$, it is obvious that $\mathrm{DS}(\hat{\gamma}_\tau)$ has a solution $\hat{\mathbf{w}}_\tau$ satisfying $\|\hat{\mathbf{w}}_\tau\|_1 \le \|\mathbf{w}^*\|_1$ for all $\tau \le r-1$. Thus (ii) also holds for $s = r-1$. Next we verify the case of $s = r$. Replacing $\mathbf{Y}_{\mathcal{I}_r}$ with $\mathbf{X}_{\mathcal{I}_r}^\top\mathbf{w}^* + \eta_{\mathcal{I}_r}$ yields the following inequality

$$\begin{aligned}
&\frac{1}{r}\left\|\mathbf{H}_{\mathcal{I}_r}\mathbf{w}^* - \hat{\mathbf{X}}_{\mathcal{I}_r}\mathbf{Y}_{\mathcal{I}_r}\right\|_\infty \\
\le &\underbrace{\frac{1}{r}\left\|\left(\mathbf{H}_{\mathcal{I}_r} - \mathbf{X}_{\mathcal{I}_r}\mathbf{X}_{\mathcal{I}_r}^\top\right)\mathbf{w}^*\right\|_\infty}_{\Xi_1} + \underbrace{\frac{1}{r}\left\|\left(\mathbf{X}_{\mathcal{I}_r} - \hat{\mathbf{X}}_{\mathcal{I}_r}\right)\mathbf{X}_{\mathcal{I}_r}^\top\mathbf{w}^*\right\|_\infty}_{\Xi_2} + \underbrace{\frac{1}{r}\left\|\hat{\mathbf{X}}_{\mathcal{I}_r}\eta_{\mathcal{I}_r}\right\|_\infty}_{\Xi_3} \\
= &\frac{1}{r}\left\|\sum_{\tau=1}^{r}\left[\left(\mathbf{H}_{\tau^2} - \mathbf{X}_{\tau^2}\mathbf{X}_{\tau^2}^\top\right)\hat{\mathbf{w}}_{\tau-1} + \left(\mathbf{H}_{\tau^2} - \mathbf{X}_{\tau^2}\mathbf{X}_{\tau^2}^\top\right)(\mathbf{w}^* - \hat{\mathbf{w}}_{\tau-1})\right]\right\|_\infty + \Xi_2 + \Xi_3 \\
\le &\underbrace{\frac{1}{r}\left\|\sum_{\tau=1}^{r}\left(\mathbf{h}_{\tau^2} - \mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top\right)\hat{\mathbf{w}}_{\tau-1}\right\|_\infty}_{\Xi_{1,1}} + \underbrace{\frac{1}{r}\left\|\sum_{\tau=1}^{r}\left(\mathbf{h}_{\tau^2} - \mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top\right)(\mathbf{w}^* - \hat{\mathbf{w}}_{\tau-1})\right\|_\infty}_{\Xi_{1,2}} + \Xi_2 + \Xi_3.
\end{aligned}$$

We separately give an upper bound on Ξ_1, Ξ_2 and Ξ_3. The key our analysis is to prove a tighter upper bound on Ξ_1 and Ξ_3, compared to the analysis in Kale et al. (2017). Ξ_1 can be decomposed into $\Xi_{1,1}$ and $\Xi_{1,2}$. By our algorithm-dependent sampling scheme, it is possible to give a tight upper bound on $\Xi_{1,1}$. The tighter upper bound of Ξ_3 comes from a more subtle analysis.

Analyzing $\Xi_{1,1}$. We define a random vector $\mathbf{z}_{\tau^2}$ as follows

$$\mathbf{z}_{\tau^2} := \mathbf{h}_{\tau^2}\hat{\mathbf{w}}_{\tau-1} - \mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top\hat{\mathbf{w}}_{\tau-1}, \quad \tau = 1,2,\ldots,r.$$

Given the selections $\{B_1,\ldots,B_{\tau-1}\}$, $\hat{\mathbf{w}}_{\tau-1}$ is deterministic. By Lemma 9, it is easy to verify that $\mathbb{E}_\tau[z_{\tau^2,i}] = 0$ for all $i \in [d]$. Therefore, $z_{1,i}, z_{2^2,i}, \ldots, z_{r^2,i}$ is a sequence of martingale differences.

By Corollary 1, the sum of the conditional variances satisfies

$$\sum_{\tau=1}^{r}\mathbb{E}_\tau[z_{\tau,i}^2]=\sum_{\tau=2}^{r}\mathbb{E}_\tau[z_{\tau,i}^2]+\mathbb{E}[z_{1,i}^2]\le\frac{2(d-1)}{k-1}\left(\sum_{\tau=2}^{r}\|\hat{\mathbf{w}}_{\tau-1}\|_1^2+\|\hat{\mathbf{w}}_0\|_1^2\right)\le 2r\frac{d-1}{k-1},$$

in which $\|\hat{\mathbf{w}}_\tau\|_1\le\|\mathbf{w}^*\|_1\le 1$ for all $\tau=1,\dots,r-1$. By Corollary 1, we have $|z_{\tau^2,i}|\le g_{d,k}$. By Lemma 6 and the union-of-events bound over $i\in[d]$, w.p. at least $1-\delta$,

$$\Xi_{1,1}\le\frac{2g_{d,k}}{3r}\ln\frac{d}{\delta}+\frac{2}{\sqrt{r}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}}.$$

Analyzing $\Xi_{1,2}$. We redefine $\mathbf{z}_{\tau^2}$ as follows

$$\mathbf{z}_{\tau^2}:=\mathbf{h}_{\tau^2}\Delta_{\tau-1}-\mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top\Delta_{\tau-1},\quad\tau=1,2,\dots,r.$$

Given the selections $\{B_1,\dots,B_{\tau-1}\}$, $\Delta_{\tau-1}$ is deterministic. It is easy to be verified that $\mathbb{E}_\tau[z_{\tau^2,i}]=0$ for all $i\in[d]$. Next we analyze the sum of conditional variances. Recalling that $\Delta_{\tau-1}(S^c)=\hat{\mathbf{w}}_\tau(S^c)$. We further decompose $z_{\tau^2,i}$ into two components

$$\underbrace{\sum_{j\in S}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j\in B_\tau]}\mathbb{I}_{i,j\in B_\tau}-x_{\tau^2,i}x_{\tau^2,j}\right)\Delta_{\tau-1,j}}_{:=z_{\tau^2,S}}+\underbrace{\sum_{j\in S^c}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j\in B_\tau]}\mathbb{I}_{i,j\in B_\tau}-x_{\tau^2,i}x_{\tau^2,j}\right)\hat{w}_{\tau-1,j}}_{:=z_{\tau^2,S^c}}.$$

Without loss of generality, assuming that $i\in S$. By Lemma 11, Lemma 12, Corollary 1 and Corollary 2, we have

$$\begin{aligned}
&\mathbb{E}_\tau[z_{\tau,i}^2]\\
=&\mathbb{E}_\tau\left[\left(z_{\tau^2,S}\right)^2\right]+\mathbb{E}_\tau\left[\left(z_{\tau^2,S^c}\right)^2\right]+2\mathbb{E}_\tau\left[z_{\tau^2,S}\cdot z_{\tau^2,S^c}\right]\\
\le&g_{d,k}\|\Delta_{\tau-1}(S)\|_1^2+\frac{2(d-1)}{k-1}\|\hat{\mathbf{w}}_{\tau-1}\|_1\cdot\|\hat{\mathbf{w}}_{\tau-1}(S^c)\|_1+\frac{2(d-1)}{k-1}\|\Delta_{\tau-1}(S)\|_1\cdot\|\hat{\mathbf{w}}_{\tau-1}(S^c)\|_1\\
\le&g_{d,k}\|\Delta_{\tau-1}(S)\|_1^2+\frac{2(d-1)}{k-1}\|\hat{\mathbf{w}}_{\tau-1}\|_1\cdot\|\hat{\mathbf{w}}_{\tau-1}(S^c)\|_1+\\
&\frac{2(d-1)}{k-1}\left(\|\mathbf{w}^*\|_1+\|\hat{\mathbf{w}}_{\tau-1}(S)\|_1\right)\cdot\left(\|\hat{\mathbf{w}}_{\tau-1}\|_1-\|\hat{\mathbf{w}}_{\tau-1}(S)\|_1\right)\\
=&g_{d,k}\|\Delta_{\tau-1}(S)\|_1^2+\frac{4(d-1)}{k-1}\|\mathbf{w}^*\|_1^2-\frac{2(d-1)}{k-1}\|\hat{\mathbf{w}}_{\tau-1}(S)\|_1\cdot\left(\|\mathbf{w}^*\|_1+\|\hat{\mathbf{w}}_{\tau-1}(S)\|_1\right)\\
\le&g_{d,k}\|\Delta_{\tau-1}(S)\|_1^2+\frac{4(d-1)}{k-1}.
\end{aligned}$$

Note that $\|\Delta_{\tau-1}(S)\|_1^2\in(0,4]$ is a random variable. By Lemma 9, we have

$$|z_{\tau^2,i}|=\left|\sum_{j=1}^{d}\left(\frac{x_{\tau^2,i}x_{\tau^2,j}}{\mathbb{P}[i,j\in B_\tau]}\mathbb{I}_{i,j\in B_\tau}-x_{\tau^2,i}x_{\tau^2,j}\right)\Delta_{\tau-1,j}\right|\le\frac{(d-1)(d-2)}{(k-1)(k-2)}\|\Delta_{\tau-1}\|_1\le 2g_{d,k}.$$

By Lemma 7 and the union-of-events bound over $i \in [d]$, w.p. at least $1 - \left(1 + \log_\beta \frac{d+k}{k-2}\right)\delta$,

$$\Xi_{1,2} \le \frac{4g_{d,k}}{3r} \ln \frac{d}{\delta} + \frac{1}{r}\sqrt{2\beta g_{d,k} \sum_{\tau=1}^{r} \|\Delta_{\tau-1}(S)\|_1^2 \ln \frac{d}{\delta}} + \sqrt{\frac{8\beta(d-1)}{r(k-1)} \ln \frac{d}{\delta}}.$$

Let $\beta = 1.5$. Combining the upper bounds on $\Xi_{1,1}$ and $\Xi_{1,2}$, w.p. at least $1 - \left(2 + \log_{1.5} \frac{d+4}{k-2}\right)\delta$,

$$\Xi_1 \le \frac{2g_{d,k}}{r} \ln \frac{d}{\delta} + \left(2 + \sqrt{12}\right)\sqrt{\frac{d-1}{r(k-1)} \ln \frac{d}{\delta}} + \frac{1}{r}\sqrt{3g_{d,k} \sum_{\tau=1}^{r} \|\Delta_{\tau-1}(S)\|_1^2 \ln \frac{d}{\delta}}.$$

Analyzing Ξ_2. Similarly, we redefine $\mathbf{z}_{\tau^2}$ as follows

$$\mathbf{z}_{\tau^2} := \mathbf{x}_{\tau^2}\hat{\mathbf{x}}_{\tau^2}^\top \mathbf{w}^* - \mathbf{x}_{\tau^2}\mathbf{x}_{\tau^2}^\top \mathbf{w}^*, \quad \tau = 1, 2, \ldots, r.$$

It is obvious $|z_{\tau^2,i}| \le \frac{d-1}{k-1}$ for all $i \in [d]$. By Lemma 9, the sum of conditional variances is upper bounded by $\frac{d-1}{k-1}r$. By Lemma 6 and the union-of-events bound over $i \in [d]$, w.p. at least $1 - \delta$,

$$\Xi_2 \le \frac{2(d-1)}{3(k-1)r} \ln \frac{d}{\delta} + \frac{1}{\sqrt{r}}\sqrt{\frac{2(d-1)}{k-1} \ln \frac{d}{\delta}}.$$

Analyzing Ξ_3. For each $i \in [d]$, we define a random variable $z_{r,i}$ as follows

$$z_{r,i} = \sum_{\tau=1}^{r} \eta_{\tau^2} \hat{x}_{\tau^2,i}, \quad i = 1, 2, \ldots, d.$$

Given that $\eta_1, \eta_4, \ldots, \eta_{r^2}$ are independent Gaussian variables, we have $\mathbb{E}_{\eta_{\tau^2}, \eta_{t^2}}[\eta_{\tau^2}\hat{x}_{\tau^2,i} \cdot \eta_{t^2}\hat{x}_{t^2,i}] = \mathbb{E}_{\eta_{t^2}}[\eta_{t^2}]\mathbb{E}_{\eta_{\tau^2}}[\eta_{\tau^2}\hat{x}_{\tau^2,i}\hat{x}_{t^2,i}] = 0$ for any $\tau \ne t$ in which we assume $\tau < t$. Thus $\eta_{\tau^2}\hat{x}_{\tau^2,i}$ is also independent of $\eta_{t^2}\hat{x}_{t^2,i}$, and $z_{r,i}$ is a Gaussian random variable with $\mathbb{E}[z_{r,i}] = 0$ and

$$\mathbb{E}[(z_{r,i})^2] = \sum_{\tau=1}^{r} \mathbb{E}[\eta_{\tau^2}^2]\hat{x}_{\tau^2,i}^2 = \sigma^2 \cdot \sum_{\tau=1}^{r} \hat{x}_{\tau^2,i}^2 \le \frac{d-1}{k-1}\sigma^2 r + \sigma^2 \cdot \sum_{\tau=1}^{r} \underbrace{\left(\hat{x}_{\tau^2,i}^2 - \frac{x_{\tau^2,i}^2}{\mathbb{P}[i \in B_\tau]}\right)}_{:=a_{\tau,i}}.$$

It can be proved that $|a_{\tau,i}| \le \frac{(d-1)^2}{(k-1)^2}$ and $\sum_{\tau=1}^{s} \mathbb{E}[a_{\tau,i}^2] \le \frac{(d-1)^3}{(k-1)^3}s$. By Lemma 6, w.p. at least $1 - \delta$,

$$\mathbb{E}[(z_{r,i})^2] \le \frac{d-1}{k-1}\sigma^2 r + \frac{2(d-1)^2}{3(k-1)^2}\sigma^2 \ln \frac{1}{\delta} + \sigma^2\sqrt{\frac{2r(d-1)^3}{(k-1)^3} \ln \frac{1}{\delta}}.$$

Denote by $\Sigma_{r,i}$ the standard variance of $z_{r,i}$. For Gaussian random variables, we have

$$\begin{aligned}
\forall z_0 > 0,\ \mathbb{P}[|z_{r,i}| > z_0] =& 2\int_{z_0}^{\infty} \frac{1}{\sqrt{2\pi}\Sigma_{r,i}} \exp\left(-\frac{z_{r,i}^2}{2\Sigma_{r,i}^2}\right) \mathrm{d}\, z_{r,i} \\
=& -\frac{2\Sigma_{r,i}}{\sqrt{2\pi} z_{r,i}} \exp\left(-\frac{z_{r,i}^2}{2\Sigma_{r,i}^2}\right)|_{z_0}^{\infty} - 2\int_{z_0}^{\infty} \frac{\Sigma_{r,i}}{\sqrt{2\pi} z_{r,i}^2} \exp\left(-\frac{z_{r,i}^2}{2\Sigma_{r,i}^2}\right) \mathrm{d}\, z_{r,i} \\
\le& 2\frac{\Sigma_{r,i}}{\sqrt{2\pi} z_0} \exp\left(-\frac{z_0^2}{2\Sigma_{r,i}^2}\right).
\end{aligned}$$

For any $\delta \in (0,1)$, let $z_0 = \Sigma_{r,i}\sqrt{2\ln\frac{1}{\delta}}$. Then we have

$$\mathbb{P}[|z_{r,i}| > z_0] \leq \frac{1}{\sqrt{\pi}\sqrt{\ln\frac{1}{\delta}}}\delta < \delta.$$

By the union-of-events bound over $i \in [d]$, with probability at least $1-2\delta$,

$$\begin{aligned}
\Xi_3 \leq & \frac{1}{r}\sqrt{\frac{d-1}{k-1}r + \frac{2(d-1)^2}{3(k-1)^2}\ln\frac{1}{\delta} + \sqrt{\frac{2r(d-1)^3}{(k-1)^3}\ln\frac{1}{\delta}}} \cdot \sigma\sqrt{2\ln\frac{d}{\delta}} \\
\leq & \frac{1}{r}\sqrt{\frac{d-1}{k-1}r + \frac{2(d-1)^2}{3(k-1)^2}\ln\frac{1}{\delta} + \frac{1}{2}\left(\frac{d-1}{k-1}\cdot\frac{3r}{4} + \frac{8(d-1)^2}{3(k-1)^2}\ln\frac{d}{\delta}\right)} \cdot \sigma\sqrt{2\ln\frac{d}{\delta}} \\
\leq & \frac{1.2\sigma}{\sqrt{r}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}} + \frac{2\sigma}{r}\frac{d-1}{k-1}\ln\frac{d}{\delta},
\end{aligned}$$

where we use the inequality $2\sqrt{ab} \leq a+b$ for $a>0, b>0$ to simplify the term in the square root. **Combining the upper bounds on** $\Xi_{1,1}, \Xi_{1,2}, \Xi_2$ **and** Ξ_3, w.p. at least $1 - \left(5 + \log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$,

$$\begin{aligned}
& \frac{1}{r}\left\|\mathbf{H}_{\mathcal{I}_r}\mathbf{w}^* - \hat{\mathbf{X}}_{\mathcal{I}_r}\mathbf{Y}_{\mathcal{I}_r}\right\|_\infty \\
\leq & \frac{2g_{d,k}}{r}\ln\frac{d}{\delta} + \left(2+\sqrt{12}\right)\sqrt{\frac{d-1}{r(k-1)}\ln\frac{d}{\delta}} + \frac{1}{r}\sqrt{3g_{d,k}\sum_{\tau=1}^{r}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}} + \\
& \frac{2(d-1)}{3(k-1)r}\ln\frac{d}{\delta} + \frac{1}{\sqrt{r}}\sqrt{\frac{2(d-1)}{k-1}\ln\frac{d}{\delta}} + \frac{1.2\sigma}{\sqrt{r}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}} + \frac{2\sigma}{r}\frac{d-1}{k-1}\ln\frac{d}{\delta} \\
\leq & \left(\frac{8}{3}+2\sigma\right)\frac{g_{d,k}}{r}\ln\frac{d}{\delta} + (6.9+1.2\sigma)\sqrt{\frac{d-1}{r(k-1)}\ln\frac{d}{\delta}} + \frac{1}{r}\sqrt{3g_{d,k}\sum_{\tau=1}^{r}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}} = \gamma_r.
\end{aligned}$$

Taking the union-of-events bound over $\tau \leq r-1$ and $\tau = r$, w.p. at least $1-r\left(5+\log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$, (i) holds for $s=r$. Since $\gamma_r \leq \hat{\gamma}_r$, $\mathrm{DS}(\hat{\gamma}_r)$ has a solution $\hat{\mathbf{w}}_r$ satisfying $\|\hat{\mathbf{w}}_r\|_1 \leq \|\mathbf{w}^*\|_1$. Thus (ii) also holds for $s=r$.

Therefore, the lemma holds for all $s \geq 1$, concluding the proof. ■

Lemma 14 *For any $s \geq 1$, with probability at least $1-\delta$,*

$$\frac{1}{s}\left\|\left(\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top - \mathbf{H}_{\mathcal{I}_s}\right)\Delta_s\right\|_\infty \leq \frac{2g_{d,k}}{3s}\|\Delta_s\|_1\ln\frac{d^2}{\delta} + \|\Delta_s\|_1\sqrt{\frac{2g_{d,k}}{s}\cdot\ln\frac{d^2}{\delta}}.$$

Proof [of Lemma 14] As Δ_s is not independent with $\mathbf{h}_{\tau^2}$, $\tau = 1,2,\ldots,s$, it is necessary to give the element-wise bound for the matrix $\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top - \mathbf{H}_{\mathcal{I}_s}$. We first analyze the elements on the diagonal. Let $z_{\tau,i} = \mathbf{h}_{\tau^2}[i,i] - \mathbf{x}_{\tau^2,i}^2$. It is obvious that

$$\forall i \in [d], \quad \mathbb{E}_\tau[z_{\tau,i}] = 0, \quad |z_{\tau,i}| \leq \frac{d-1}{k-1}, \quad \sum_{\tau=1}^{s}\mathbb{E}_\tau[|z_{\tau,i}|^2] \leq \sum_{\tau=1}^{s}\frac{x_{\tau^2,i}^4}{\mathbb{P}[i\in B_\tau]} \leq \frac{d-1}{k-1}s.$$

By Lemma 6, with probability at least $1-\delta$,

$$\left|\sum_{\tau=1}^{s} z_{\tau,i}\right| \le \frac{2(d-1)}{3(k-1)}\ln\frac{2}{\delta} + \sqrt{2s\frac{d-1}{k-1}\ln\frac{2}{\delta}}.$$

Next we analyze the non-diagonal elements. Let $z_{\tau,i,j} = \mathbf{h}_{\tau^2}[i,j] - \mathbf{x}_{\tau^2,i}\mathbf{x}_{\tau^2,j}$. We have

$$\forall i \neq j \in [d], \quad \mathbb{E}_\tau[z_{\tau,i,j}] = 0, \quad |z_{\tau,i,j}| \le g_{d,k}, \quad \sum_{\tau=1}^{s}\mathbb{E}_\tau[|z_{\tau,i,j}|^2] = \sum_{\tau=1}^{s}\frac{x_{\tau^2,i}^2 x_{\tau^2,j}^2}{\mathbb{P}[i,j\in B_\tau]} \le g_{d,k}\cdot s.$$

By Lemma 6, with probability at least $1-\delta$,

$$\left|\sum_{\tau=1}^{s} z_{\tau,i,j}\right| \le \frac{2(d-1)(d-2)}{3(k-1)(k-2)}\ln\frac{2}{\delta} + \sqrt{2s\frac{(d-1)(d-2)}{(k-1)(k-2)}\ln\frac{2}{\delta}}.$$

Since the matrix is symmetry, it is enough to take the union-of-events bound over $d+\frac{d(d-1)}{2}$ events. With probability at least $1-\delta$,

$$\begin{aligned}
&\frac{1}{s}\left\|\left(\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top - \mathbf{H}_{\mathcal{I}_s}\right)\Delta_s\right\|_\infty \\
\le&\frac{1}{s}\max_{i\in[d]}\left|\sum_{\tau=1}^{s} z_{\tau,i}\Delta_i + \sum_{\tau=1}^{s}\sum_{j\neq i} z_{\tau,i,j}\Delta_j\right| \\
\le&\frac{1}{s}\max_{i\in[d]}\left(\left|\sum_{\tau=1}^{s} z_{\tau,i}\right|\cdot|\Delta_i| + \sum_{j\neq i}\left|\sum_{\tau=1}^{s} z_{\tau,i,j}\right|\cdot|\Delta_j|\right) \\
\le&\frac{1}{s}\max_{i\in[d]}\left(\frac{2(d-1)(d-2)}{3(k-1)(k-2)}\ln\frac{d(d+1)}{\delta} + \sqrt{2s\frac{(d-1)(d-2)}{(k-1)(k-2)}\ln\frac{d(d+1)}{\delta}}\right)\sum_{i=1}^{j}|\Delta_i| \\
\le&\frac{2g_{d,k}}{3s}\|\Delta_s\|_1\ln\frac{d(d+1)}{\delta} + \|\Delta_s\|_1\sqrt{\frac{2g_{d,k}}{s}\cdot\ln\frac{d(d+1)}{\delta}},
\end{aligned}$$

which concludes the proof. ■

Lemma 15 *For any $s \ge 1$, if $\hat{\gamma}_\tau \ge \gamma_\tau$ for all $\tau \le s$, then w.p. at least $1 - s\left(6+\log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$,*

$$\begin{aligned}
\forall \tau \le s, \quad \left\|\frac{1}{\tau}\mathbf{X}_\tau\mathbf{X}_\tau^\top\Delta_\tau\right\|_\infty \le& 2\left(\frac{8}{3}+2\sigma\right)\frac{g_{d,k}}{\tau}\ln\frac{d}{\delta} + 2\left(6.9+1.2\sigma\right)\sqrt{\frac{d-1}{\tau(k-1)}\ln\frac{d}{\delta}} + \nu_\tau + \\
&\frac{1}{\tau}\sqrt{3g_{d,k}\sum_{r=1}^{\tau}\|\Delta_{r-1}(S)\|_1^2\ln\frac{d}{\delta}} + \frac{2g_{d,k}}{3\tau}\|\Delta_\tau\|_1\ln\frac{d^2}{\delta} + \sqrt{\frac{2g_{d,k}}{\tau}\ln\frac{d^2}{\delta}}\|\Delta_\tau\|_1.
\end{aligned}$$

Proof [of Lemma 15] We decompose $\left\|\frac{1}{\tau}\mathbf{X}_{\mathcal{I}_\tau}\mathbf{X}_{\mathcal{I}_\tau}^\top\Delta_\tau\right\|_\infty$ into three components.

$$\begin{aligned}
&\left\|\frac{1}{\tau}\mathbf{X}_{\mathcal{I}_\tau}\mathbf{X}_{\mathcal{I}_\tau}^\top\Delta_\tau\right\|_\infty \\
=&\left\|\frac{1}{\tau}\mathbf{H}_{\mathcal{I}_\tau}\Delta_\tau\right\|_\infty+\left\|\frac{1}{\tau}\left(\mathbf{X}_{\mathcal{I}_\tau}\mathbf{X}_{\mathcal{I}_\tau}^\top-\mathbf{H}_{\mathcal{I}_\tau}\right)\Delta_\tau\right\|_\infty \\
=&\frac{1}{\tau}\left\|\mathbf{H}_{\mathcal{I}_\tau}\hat{\mathbf{w}}_\tau-\hat{\mathbf{X}}_{\mathcal{I}_\tau}^\top\mathbf{Y}_{\mathcal{I}_\tau}-\left(\mathbf{H}_{\mathcal{I}_\tau}\mathbf{w}^*-\hat{\mathbf{X}}_{\mathcal{I}_\tau}^\top\mathbf{Y}_{\mathcal{I}_\tau}\right)\right\|_\infty+\left\|\frac{1}{\tau}\left(\mathbf{X}_{\mathcal{I}_\tau}\mathbf{X}_{\mathcal{I}_\tau}^\top-\mathbf{H}_{\mathcal{I}_\tau}\right)\Delta_\tau\right\|_\infty \\
\leq&\frac{1}{\tau}\left\|\mathbf{H}_{\mathcal{I}_\tau}\hat{\mathbf{w}}_\tau-\hat{\mathbf{X}}_{\mathcal{I}_\tau}^\top\mathbf{Y}_{\mathcal{I}_\tau}\right\|_\infty+\frac{1}{\tau}\left\|\mathbf{H}_{\mathcal{I}_\tau}\mathbf{w}^*-\hat{\mathbf{X}}_{\mathcal{I}_\tau}^\top\mathbf{Y}_{\mathcal{I}_\tau}\right\|_\infty+\left\|\frac{1}{\tau}\left(\mathbf{X}_{\mathcal{I}_\tau}\mathbf{X}_{\mathcal{I}_\tau}^\top-\mathbf{H}_{\mathcal{I}_\tau}\right)\Delta_\tau\right\|_\infty.
\end{aligned}$$

If $\hat{\gamma}_\tau \geq \gamma_\tau$ for all $\tau \leq s$, then Lemma 13 ensures that the second component is upper bounded by γ_τ, and the first component is upper bounded by $\hat{\gamma}_\tau$. By Lemma 14, we can also obtain an upper bound on the third component. Combining all results concludes the proof. ■

Lemma 16 *Let $S = \mathrm{Supp}(\mathbf{w}^*)$. For any $s \geq 1$, if $\hat{\gamma}_\tau \geq \gamma_\tau$ for all $\tau \leq s$, then w.p. at least $1 - s\left(5 + \log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$,*

$$\forall \tau \leq s, \quad \|\Delta_\tau(S^c)\|_1 \leq \|\Delta_\tau(S)\|_1,$$

in which $\Delta_\tau = \hat{\mathbf{w}}_\tau - \mathbf{w}^*$ and $\hat{\mathbf{w}}_\tau$ be the solution of $\mathrm{Ds}(\hat{\gamma}_\tau)$.

Recalling the definition of the restricted set Ω_S in Assumption 3, Δ_τ belongs to Ω_S with $\alpha = 1$. In this way, it is possible to use the (δ_S, S, α)-compatibility condition.

Proof [of Lemma 16] By Lemma 13, we obtain, w.p. at least $1 - s\left(5 + \log_{1.5}\frac{2(d-2)}{3(k-2)}\right)\delta$, $\|\hat{\mathbf{w}}_\tau\|_1 \leq \|\mathbf{w}^*\|_1$ for all $\tau \leq s$. Since $w_i^* = 0$ for $i \in S^c$. We have

$$\|\Delta_\tau(S^c)\|_1 = \|\hat{\mathbf{w}}_\tau(S^c)\|_1 = \|\hat{\mathbf{w}}_\tau\|_1 - \|\hat{\mathbf{w}}_\tau(S)\|_1 \leq \|\mathbf{w}^*(S)\|_1 - \|\hat{\mathbf{w}}_\tau(S)\|_1 \leq \|\mathbf{w}^*(S) - \hat{\mathbf{w}}_\tau(S)\|_1,$$

which concludes the proof. ■

Appendix C. Proof of Lemma 2

Proof [of Lemma 2] Let $s' \in [1, \lfloor\sqrt{T}\rfloor]$. For any $s \in [1, s']$, we will separately give a lower bound and an upper bound on $\frac{1}{s}\|\mathbf{X}_{\mathcal{I}_s}^\top\Delta_s\|_2^2$. If $\hat{\gamma}_s \geq \gamma_s$ for all $1 \leq s \leq s'$, then by Lemma 16 and Assumption 3, we have, w.p. at least $1 - s'\left(5 + \log_{1.5}\frac{d+k}{k-2}\right)\delta$, a lower bound is given as follows,

$$\forall s \in [1, s'], \quad \frac{1}{s}\|\mathbf{X}_{\mathcal{I}_s}^\top\Delta_s\|_2^2 \geq \frac{\delta_S^2}{|S|}\|\Delta_s(S)\|_1^2 \geq \frac{\delta_S^2}{k}\|\Delta_s(S)\|_1^2,$$

in which $|S| = \|\mathbf{w}^*\|_0 \leq k$. Besides, an upper bound is as follows,

$$\forall s \in [1, s'], \quad \frac{1}{s}\|\mathbf{X}_{\mathcal{I}_s}^\top\Delta_s\|_2^2 \leq \left\|\frac{1}{s}\Delta_s^\top\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top\right\|_\infty\|\Delta_s\|_1 \leq 2\left\|\frac{1}{s}\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top\Delta_s\right\|_\infty\|\Delta_s(S)\|_1.$$

Lemma 15 further provides an upper bound on $\left\|\frac{1}{s}\mathbf{X}_{\mathcal{I}_s}\mathbf{X}_{\mathcal{I}_s}^\top\Delta_s\right\|_\infty$. Combining the lower bound and upper bound, we obtain, w.p. at least $1-s'\left(6+\log_{1.5}\frac{d+k}{k-2}\right)\delta$,

$$\forall s\le s',\quad \frac{\delta_S^2\|\Delta_s(S)\|_1}{2k}\le 2\left(\frac{8}{3}+2\sigma\right)\frac{g_{d,k}}{s}\ln\frac{d}{\delta}+2\left(6.9+1.2\sigma\right)\sqrt{\frac{d-1}{s(k-1)}\ln\frac{d}{\delta}}+\nu_s+\frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}}+\frac{2g_{d,k}}{3s}\|\Delta_s\|_1\ln\frac{d^2}{\delta}+\sqrt{\frac{2g_{d,k}}{s}\ln\frac{d^2}{\delta}}\|\Delta_s\|_1. \tag{16}$$

Solving the inequality (16) will provide an explicit upper bound on $\|\Delta_s(S)\|_1$. However, it is highly non-trivial, as the right-hand side of the inequality depends on $\sqrt{\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2}$ and ν_s that will be estimated dynamically. An obvious upper bound can be derived by using the inequality $\|\Delta_\tau(S)\|_1\le 2$ for all $\tau\le s-1$. However, such a simple analysis overlooked the fact that $\|\Delta_\tau(S)\|_1$ becomes smaller as τ increases. We will carefully control $\sqrt{\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2}$ by an induction method, naturally yielding a tighter upper bound of $\|\Delta_s(S)\|_1$. Next we consider three cases. **It is crucial to verify $\hat{\gamma}_s\ge\gamma_s$ for all $s\in[1,s']$.**

C.1. Case 1: $s\le s_0$

Let $s'=s_0$. We use the trivial upper bound $\|\Delta_\tau(S)\|_1\le 2$. Recalling that ν_s is defined as follows,

$$\forall s\le s_0,\quad \nu_s=\frac{2}{\sqrt{s}}\sqrt{3g_{d,k}\ln\frac{d}{\delta}}\ge\frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}}\Rightarrow\hat{\gamma}_s\ge\gamma_s.$$

Therefore, for any $s\le s_0$, **by Lemma 13, w.p. at least** $1-s_0\left(6+\log_{1.5}\frac{d+k}{k-2}\right)\delta$, $\mathrm{DS}(\hat{\gamma}_s)$ **has a feasible solution at least**, i.e., $\mathbf{w}^*$. It is proper to derive the inequality (16). We further obtain

$$\frac{\delta_S^2\|\Delta_s(S)\|_1}{2k}\le\left(8+4\sigma\right)\frac{g_{d,k}}{s}\ln\frac{d}{\delta}+2\left(6.9+1.2\sigma\right)\sqrt{\frac{(d-1)\ln\frac{d}{\delta}}{s(k-1)}}+\frac{4+4\sqrt{3}}{\sqrt{s}}\sqrt{g_{d,k}\ln\frac{d}{\delta}},$$

where we use the inequality $\ln\frac{d^2}{\delta}\le 2\ln\frac{d}{\delta}$. Rearranging terms concludes the desired result.

C.2. Case 2: $s_0<s\le s_1$

Let $s'=s_1$. We start from (16). Substituting into the value of s_0, the inequality (17) holds.

$$\forall s>s_0,\quad 4k\sqrt{\frac{2g_{d,k}}{s}\cdot\ln\frac{d^2}{\delta}}\le\frac{1}{4}\delta_S^2. \tag{17}$$

If $\hat{\gamma}_s\ge\gamma_s$ for all $s\in(s_0,s_1]$, then, the inequality (16) can be rewritten as follows.

$$\frac{\delta_S^2\|\Delta_s(S)\|_1}{2k}\le\left(8+4\sigma\right)\frac{g_{d,k}}{s}\ln\frac{d}{\delta}+\frac{2(6.9+1.2\sigma)}{\sqrt{s}}\sqrt{\frac{d-1}{k-1}\ln\frac{d}{\delta}}+\nu_s+\frac{1}{s}\sqrt{3g_{d,k}\sum_{\tau=1}^{s}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}}+\frac{\delta_S^2\|\Delta_s(S)\|_1}{8k},$$

in which we use Lemma 16 and (17). Rearranging terms and substituting into the value of s_0 yields,

$$\begin{aligned}
&\|\Delta_s(S)\|_1\\
\leq&\frac{8k\nu_s}{3\delta_S^2}+\frac{a_1kg_{d,k}}{\delta_S^2 s}+\frac{a_2k}{\delta_S^2\sqrt{s}}\sqrt{\frac{d-1}{k-1}}+\frac{a_3k\sqrt{g_{d,k}}}{\delta_S^2 s}\sqrt{\sum_{\tau=1}^{s_0}\|\Delta_{\tau-1}(S)\|_1^2+\sum_{\tau=s_0+1}^{s}\|\Delta_{\tau-1}(S)\|_1^2}\\
\leq&\frac{8k\nu_s}{3\delta_S^2}+\frac{\delta_S^2\frac{a_1}{k}+48a_3\sqrt{\ln\frac{d^2}{\delta}}}{\delta_S^4}\frac{k^2g_{d,k}}{s}+\frac{a_2k}{\delta_S^2\sqrt{s}}\sqrt{\frac{d-1}{k-1}}+\frac{a_3k\sqrt{g_{d,k}}}{s\delta_S^2}\sqrt{\sum_{\tau=s_0}^{s-1}\|\Delta_\tau(S)\|_1^2},
\end{aligned}\tag{18}$$

in which a_1, a_2 and a_3 are defined in Lemma 2. For simplicity, let

$$a_4'=\delta_S^2\frac{a_1}{k}+48a_3\sqrt{\ln\frac{d^2}{\delta}}.$$

Next we will use the induction method to prove the convergence rate of $\|\Delta_s(S)\|_1$. To be specific, we will prove that, w.p. at least $1-s_1\left(6+\log_{1.5}\frac{d+k}{k-2}\right)\delta$,

$$\begin{aligned}
\forall s_0<s\leq s_1,\quad \|\Delta_s(S)\|_1\leq&\frac{a_4}{1-\frac{2\sqrt{3}}{9}}\frac{k^2g_{d,k}}{s\delta_S^4},\\
a_4=&\delta_S^2\frac{a_1}{k}+24a_2\sqrt{\frac{k-2}{d-2}\ln\frac{d^2}{\delta}}+4a_3\left(24\sqrt{\ln\frac{d^2}{\delta}}+\frac{\delta_S^2}{k\sqrt{g_{d,k}}}\right).
\end{aligned}\tag{19}$$

We first analyze the case $s=s_0+1$. Recalling that $\mu_1=\frac{9}{9-2\sqrt{3}}$ and

$$\begin{aligned}
\nu_{s_0+1}=&\frac{24kg_{d,k}}{(s_0+1)\delta_S^2}\sqrt{12\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}}+\frac{2\sqrt{3g_{d,k}\ln\frac{d}{\delta}}}{s_0+1}\geq\frac{1}{s_0+1}\sqrt{3g_{d,k}\sum_{\tau=1}^{s_0+1}\|\Delta_{\tau-1}(S)\|_1^2\ln\frac{d}{\delta}},\\
\nu_s=&\frac{24kg_{d,k}}{s\delta_S^2}\sqrt{12\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}}+\frac{2\sqrt{3g_{d,k}\ln\frac{d}{\delta}}}{s}+\frac{\mu_1a_4\sqrt{3g_{d,k}}}{\delta_S^4\cdot s}\sqrt{\sum_{\tau=s_0+1}^{s-1}\frac{k^4g_{d,k}^2}{\tau^2}\ln\frac{d}{\delta}},\\
&\forall s_0+1<s\leq s_1.
\end{aligned}$$

We still have $\hat{\gamma}_{s_0+1}\geq\gamma_{s_0+1}$. By Lemma 13, with a high probability, $\mathrm{DS}(\hat{\gamma}_{s_0+1})$ has a feasible solution at least. In this way, it is proper to derive (18) for any $s\leq s_0+1$.

$$\begin{aligned}
&\|\Delta_{s_0+1}(S)\|_1\\
\leq&\frac{8k\nu_{s_0+1}}{3\delta_S^2}+a_4'\frac{k^2g_{d,k}}{(s_0+1)\delta_S^4}+\frac{a_2k}{\delta_S^2}\sqrt{\frac{d-1}{(s_0+1)(k-1)}}+\frac{a_3k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{(s_0+1)}\sqrt{\|\Delta_{s_0}(S)\|_1^2}\\
\leq&\left(64\sqrt{12\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}}+a_4'+a_2\sqrt{24^2\frac{k-2}{d-2}\ln\frac{d^2}{\delta}}+\frac{16\sqrt{3\ln\frac{d}{\delta}}\delta_S^2}{3k\sqrt{g_{d,k}}}+\frac{2a_3\delta_S^2}{k\sqrt{g_{d,k}}}\right)\frac{k^2g_{d,k}}{(s_0+1)\delta_S^4}\\
\leq&a_4\frac{k^2g_{d,k}}{(s_0+1)\delta_S^4},
\end{aligned}$$

in which

$$a_4 = \delta_S^2 \frac{a_1}{k} + 24a_2\sqrt{\frac{k-2}{d-2}\ln\frac{d^2}{\delta}} + 4a_3\left(24\sqrt{\ln\frac{d^2}{\delta}} + \frac{\delta_S^2}{k\sqrt{g_{d,k}}}\right).$$

Thus (19) holds for $s = s_0 + 1$.

Now assuming that (19) holds for any $s = r \in [s_0 + 1, s_1)$, **implying that** $\hat{\gamma}_{r+1} \geq \gamma_{r+1}$ **and** $\mathrm{DS}(\hat{\gamma}_{r+1})$ **has a feasible solution at least**. In this way, it is proper to derive (18) for any $s \leq r+1$, and more importantly, it is possible to analyze the upper bound on $\|\Delta_{r+1}(S)\|_1$.

$$\begin{aligned}
&\|\Delta_{r+1}(S)\|_1 \\
\leq&\frac{8k\nu_{r+1}}{3\delta_S^2} + \frac{a_4' k^2 g_{d,k}}{(r+1)\delta_S^4} + \frac{a_2 k}{\delta_S^2}\sqrt{\frac{d-1}{(r+1)(k-1)}} + \frac{a_3 k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{r+1}\sqrt{\|\Delta_{s_0}(S)\|_1^2 + \sum_{\tau=s_0+1}^{r}\|\Delta_\tau(S)\|_1^2} \\
\leq&a_4\frac{k^2 g_{d,k}}{(r+1)\delta_S^4} + \frac{8k\mu_1 a_4\sqrt{3g_{d,k}\ln\frac{d}{\delta}}}{3\delta_S^6(r+1)}\sqrt{\sum_{\tau=s_0+1}^{r}\frac{k^4 g_{d,k}^2}{\tau^2}} + \frac{a_3 k}{\delta_S^6}\frac{\sqrt{g_{d,k}}}{r+1}\cdot\mu_1 a_4\sqrt{\sum_{\tau=s_0+1}^{r}\frac{k^4 g_{d,k}^2}{\tau^2}} \\
\leq&a_4\frac{k^2 g_{d,k}}{(r+1)\delta_S^4} + 2a_3\frac{k}{\delta_S^6}\frac{\sqrt{g_{d,k}}}{r+1}\cdot\mu_1 a_4\sqrt{k^4 g_{d,k}^2\left(\frac{1}{s_0}-\frac{1}{r}\right)} \qquad \text{(By Lemma 4)} \\
\leq&a_4\frac{k^2 g_{d,k}}{(r+1)\delta_S^4} + 2a_3\frac{k}{\delta_S^6}\frac{\sqrt{g_{d,k}}}{r+1}\cdot\mu_1 a_4\sqrt{\frac{k^4 g_{d,k}^2}{\frac{24^2 g_{d,k}k^2}{\delta_S^4}\ln\frac{d^2}{\delta}}} \\
=&\mu_1 a_4\frac{k^2 g_{d,k}}{\delta_S^4(r+1)}.
\end{aligned}$$

Thus (19) also holds for $s = r+1$. We conclude that (19) holds for all $s \in (s_0, s_1]$.

C.3. Case 3: $s_1 < s < \lfloor\sqrt{T}\rfloor$

Let $s' = \lfloor\sqrt{T}\rfloor$. Recalling that $\mu_2 = \left(1 - \frac{\sqrt{6}}{9\sqrt{\frac{d-2}{k-2}\ln\frac{d^2}{\delta}}}\right)^{-1}$ and

$$\begin{aligned}
\nu_{s_1+1} =&\frac{\sqrt{3g_{d,k}}}{s_1+1}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d,k}\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}} + \sqrt{4\ln\frac{d}{\delta}} + \frac{\mu_1 a_4 k^2 g_{d,k}}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s_1}\frac{1}{\tau^2}\ln\frac{d}{\delta}}\right), \\
\nu_s =&\frac{\sqrt{3g_{d,k}}}{s}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d,k}\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}} + \sqrt{4\ln\frac{d}{\delta}} + \frac{\mu_1 a_4 k^2 g_{d,k}}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s_1}\frac{1}{\tau^2}\ln\frac{d}{\delta}} + \right. \\
&\left.\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\sum_{\tau=s_1+1}^{s-1}\frac{k^2(d-1)}{\tau(k-1)}\ln\frac{d}{\delta}}\right), \quad \forall s > s_1+1.
\end{aligned}$$

As (19) holds, it can be verified that $\hat{\gamma}_{s_1+1} \geq \gamma_{s_1+1}$. **By Lemma 13, with a high probability,** $\mathrm{DS}(\hat{\gamma}_{s_1+1})$ **has a feasible solution at least**. In this way, it is proper to derive (18) for any $s \leq s_1+1$.

Similarly, we also use the induction method to prove the convergence rate of $\|\Delta_s(S)\|_1$ and $\mathrm{DS}(\hat{\gamma}_s)$ has a feasible solution for all $s > s_1$. Specifically, by (18), we will prove that, with probability at least $1 - \sqrt{T}\left(6 + \log_{1.5}\frac{d+4}{k-2}\right)\delta$,

$$\begin{aligned}
&\forall s > s_1, \quad \|\Delta_s(S)\|_1 \leq \frac{\mu_2 a_5 k}{\delta_S^2}\sqrt{\frac{d-1}{s(k-1)}}, \\
&a_5 = \mu_1\left(\delta_S^2\frac{\frac{8}{9}+\frac{4\sigma}{9}}{k} + \frac{32\sqrt{3}}{3} + \frac{4\sqrt{3}\delta_S^2}{9k\sqrt{g_{d,k}\ln\frac{d^2}{\delta}}}\right) + a_2 + (\mu_1 - 1)a_2\sqrt{\frac{k-2}{(d-2)\ln\frac{d^2}{\delta}}}.
\end{aligned} \tag{20}$$

We first verify the case $s = s_1 + 1$.

$$\begin{aligned}
&\|\Delta_{s_1+1}(S)\|_1 \\
\leq&\frac{8k\nu_{s_1+1}}{3\delta_S^2} + \frac{a_4' k^2 g_{d,k}}{(s_1+1)\delta_S^4} + \frac{a_2 k}{\delta_S^2\sqrt{s_1+1}}\sqrt{\frac{d-1}{k-1}} + \frac{a_3 k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{s_1+1}\sqrt{\|\Delta_{s_0}(S)\|_1^2 + \sum_{s=s_0+1}^{s_1}\|\Delta_s(S)\|_1^2} \\
\leq&\underbrace{\frac{8k\nu_{s_1+1}}{3\delta_S^2}}_{\Xi_1} + \underbrace{\frac{a_4' k^2 g_{d,k}}{(s_1+1)\delta_S^4} + \frac{a_2 k}{\delta_S^2\sqrt{s_1+1}}\sqrt{\frac{d-1}{k-1}}}_{\Xi_2} + \underbrace{\frac{2a_3 k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{s_1+1}}_{\Xi_3} + \underbrace{\frac{a_3 k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{s_1+1}\sqrt{\sum_{s=s_0+1}^{s_1}\|\Delta_s(S)\|_1^2}}_{\Xi_4}.
\end{aligned}$$

Next we separately analyze the four terms on the right-hand side of the inequality.

First, we analyze Ξ_1. Substituting into the value of ν_{s_1+1} and s_0, we obtain

$$\begin{aligned}
&\Xi_1 \\
=&\frac{8k}{3\delta_S^2}\frac{\sqrt{3g_{d,k}}}{s_1+1}\left(\frac{48k}{\delta_S^2}\sqrt{g_{d,k}\ln\left(\frac{d}{\delta}\right)\ln\frac{d^2}{\delta}} + \sqrt{4\ln\frac{d}{\delta}} + \frac{\mu_1 a_4 k^2 g_{d,k}}{\delta_S^4}\sqrt{\sum_{\tau=s_0+1}^{s_1-1}\frac{1}{\tau^2}\ln\frac{d}{\delta}}\right) \\
\leq&\frac{1}{\delta_S^2\sqrt{s_1+1}}\left(a_3 k\frac{g_{d,k}}{\sqrt{s_1+1}}\frac{48k}{\delta_S^2}\sqrt{\ln\frac{d^2}{\delta}} + 2a_3 k\frac{\sqrt{g_{d,k}}}{\sqrt{s_1+1}} + a_3 k\frac{\sqrt{g_{d,k}}}{\sqrt{s_1+1}}\frac{\mu_1 a_4 k^2 g_{d,k}}{\delta_S^4}\sqrt{\frac{1}{s_0}}\right) \\
=&\frac{1}{\delta_S^2\sqrt{s_1+1}}\left(\frac{2a_3\sqrt{d-1}k}{\sqrt{(k-1)\ln\frac{d}{\delta}}} + \frac{a_3\delta_S^2}{12k\sqrt{g_{d,k}\ln(\frac{d}{\delta})\ln\frac{d^2}{\delta}}}k\sqrt{\frac{d-1}{k-1}} + \frac{a_3\mu_1 a_4}{24^2\sqrt{\ln\frac{d}{\delta}\ln\frac{d^2}{\delta}}}\frac{k\sqrt{d-1}}{\sqrt{k-1}}\right) \\
=&\frac{k}{\delta_S^2}\frac{\sqrt{d-1}}{\sqrt{(k-1)(s_1+1)}}\left(\frac{2a_3}{\sqrt{\ln\frac{d}{\delta}}} + \frac{a_3\delta_S^2}{12k\sqrt{g_{d,k}\ln(\frac{d}{\delta})\ln\frac{d^2}{\delta}}} + \frac{a_3\mu_1 a_4}{24^2\sqrt{\ln\frac{d}{\delta}\ln\frac{d^2}{\delta}}}\right),
\end{aligned}$$

in which $a_3 = \frac{8\sqrt{3}}{3}\sqrt{\ln\frac{d}{\delta}}$ and the first inequality comes from Lemma 4.

Next we analyze Ξ_2. Substituting into the value of s_1, a_1, a_3 and a_4' yields

$$\begin{aligned}
\Xi_2 \leq & \left(a_4' \frac{k g_{d,k}}{\sqrt{\frac{24^2 g_{d,k} k^2}{\delta_S^4} \frac{d-2}{k-2} \ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta} \delta_S^2}} \frac{\sqrt{k-1}}{\sqrt{d-1}} + a_2 \right) \frac{k}{\delta_S^2} \sqrt{\frac{d-1}{(s_1+1)(k-1)}} \\
\leq & \left(\frac{a_4'}{24\sqrt{\ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta}}} + a_2 \right) \frac{k}{\delta_S^2} \sqrt{\frac{d-1}{(s_1+1)(k-1)}} \\
\leq & \left(\frac{\delta_S^2 \frac{1}{k} (\frac{64}{3} + \frac{32}{3}\sigma) \ln \frac{d}{\delta} + 48 a_3 \sqrt{\ln \frac{d^2}{\delta}}}{24\sqrt{\ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta}}} + a_2 \right) \frac{k}{\delta_S^2} \sqrt{\frac{d-1}{(s_1+1)(k-1)}} \\
\leq & \left(\delta_S^2 \frac{\frac{8}{9} + \frac{4\sigma}{9}}{k} + \frac{2a_3}{\sqrt{\ln \frac{d}{\delta}}} + a_2 \right) \frac{k}{\delta_S^2} \sqrt{\frac{d-1}{(s_1+1)(k-1)}}.
\end{aligned}$$

For Ξ_3, we can obtain

$$\Xi_3 = \frac{2a_3 k \sqrt{g_{d,k}}}{\sqrt{\frac{24^2 g_{d,k} k^2}{\delta_S^4} \frac{d-2}{k-2} \ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta}}} \frac{1}{\delta_S^2 \sqrt{s_1+1}} = \frac{a_3 \delta_S^2}{12k\sqrt{g_{d,k} \ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta}}} \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}}.$$

Finally, we analyze Ξ_4. By Lemma 4 and (19), we can obtain

$$\begin{aligned}
\frac{a_3 k}{\delta_S^2} \frac{\sqrt{g_{d,k}}}{s_1+1} \sqrt{\sum_{s=s_0+1}^{s_1} \|\Delta_s(S)\|_1^2} \leq & \frac{a_3}{24k\sqrt{g_{d,k} \ln\left(\frac{d}{\delta}\right) \ln \frac{d^2}{\delta}}} \frac{k\sqrt{d-1}}{\sqrt{(s_1+1)(k-1)}} \frac{\mu_1 a_4}{\delta_S^4} \cdot \sqrt{\frac{k^4 g_{d,k}^2}{s_0}} \\
= & \frac{a_3 \mu_1 a_4}{24^2 \sqrt{\ln \frac{d}{\delta} \ln \frac{d^2}{\delta}}} \cdot \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}}.
\end{aligned}$$

Combining the upper bounds on Ξ_1, Ξ_2, Ξ_3 and Ξ_4, we obtain

$$\begin{aligned}
& \|\Delta_{s_1+1}(S)\|_1 \\
\leq & \left(\frac{4a_3}{\sqrt{\ln \frac{d}{\delta}}} + \frac{2a_3 \delta_S^2}{12k\sqrt{g_{d,k} \ln(\frac{d}{\delta}) \ln \frac{d^2}{\delta}}} + \frac{2a_3 \mu_1 a_4}{24^2 \sqrt{\ln \frac{d}{\delta} \ln \frac{d^2}{\delta}}} + \delta_S^2 \frac{\frac{8}{9} + \frac{4\sigma}{9}}{k} + a_2 \right) \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}} \\
\leq & \left(\delta_S^2 \frac{\frac{8}{9} + \frac{4\sigma}{9}}{k} + \frac{32\sqrt{3}}{3} + \frac{4\sqrt{3}\delta_S^2}{9k\sqrt{g_{d,k} \ln \frac{d^2}{\delta}}} + a_2 \right) \cdot \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}} + \\
& (\mu_1 - 1) \left(\delta_S^2 \frac{\frac{8}{9} + \frac{4\sigma}{9}}{k} + a_2 \sqrt{\frac{k-2}{(d-2)\ln \frac{d^2}{\delta}}} + \frac{4a_3}{\sqrt{\ln \frac{d^2}{\delta}}} + \frac{4a_3 \delta_S^2}{24k\sqrt{g_{d,k} \ln \frac{d^2}{\delta}}} \right) \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}} \\
\leq & a_5 \frac{k\sqrt{d-1}}{\delta_S^2 \sqrt{(s_1+1)(k-1)}}.
\end{aligned}$$

Thus (20) holds for $s = s_1 + 1$.

Now assuming that (20) holds for any $s = r \geq s_1 + 1$, **implying that** $\hat{\gamma}_{r+1} \geq \gamma_{r+1}$ **and** $\mathrm{DS}(\hat{\gamma}_{r+1})$ **has a feasible solution**. In this way, it is proper to derive (18) for any $s \leq r + 1$, By the second inequality in Lemma 4, we obtain

$$
\begin{aligned}
&\|\Delta_{r+1}(S)\|_1 \\
\leq&\frac{8k\nu_{r+1}}{3\delta_S^2} + \frac{a_4' k^2 g_{d,k}}{(r+1)\delta_S^4} + \frac{a_2 k}{\delta_S^2\sqrt{r+1}}\sqrt{\frac{d-1}{k-1}} + \frac{a_3}{\delta_S^2}\frac{k\sqrt{g_{d,k}}}{r+1}\sqrt{\sum_{s=s_0}^{s_1}\|\Delta_s(S)\|_1^2 + \sum_{\tau=s_1+1}^{r}\|\Delta_\tau(S)\|_1^2} \\
\leq&\frac{a_5 k}{\delta_S^2}\sqrt{\frac{d-1}{(k-1)(r+1)}} + 2a_3\frac{k}{\delta_S^2}\frac{\sqrt{g_{d,k}}}{\sqrt{r+1}}\cdot\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\sum_{s=s_1+1}^{r}\frac{k^2(d-1)}{(r+1)s(k-1)}} \\
\leq&\frac{a_5 k}{\delta_S^2}\sqrt{\frac{d-1}{(k-1)(r+1)}} + 2a_3\frac{k\sqrt{g_{d,k}}}{\delta_S^2\sqrt{2s_1}}\cdot\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\frac{k^2(d-1)}{(r+1)(k-1)}} \\
=&\frac{a_5 k}{\delta_S^2}\sqrt{\frac{d-1}{(k-1)(r+1)}} + \frac{\sqrt{6}}{9\sqrt{\frac{d-2}{k-2}\ln\frac{d^2}{\delta}}}\cdot\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\frac{k^2(d-1)}{2(r+1)(k-1)}} \\
=&\frac{\mu_2 a_5}{\delta_S^2}\sqrt{\frac{k(d-1)}{(r+1)(k-1)}}.
\end{aligned}
$$

Thus (20) also holds for $s = r + 1$. Therefore, (20) holds for all $s > s_1$, concluding the proof. ■

Appendix D. Proof of Theorem 1

Lemma 17 *Let*

$$s_2 = 4\cdot\frac{(\mu_2 a_5)^2}{\delta_S^4}\cdot\frac{k^2(d-1)}{\min_{i\in S}|w_i^*|^2\cdot(k-1)}.$$

For any $s > s_2$, with probability at least $1 - \sqrt{T}\left(6 + \log_{1.5}\frac{d+k}{k-2}\right)\delta$, it must be $S_s = S$.

Proof [of Lemma 17] Assuming that $S_s \neq S$, then there is a $j \in S_s \setminus S$ and $i \in S \setminus S_s$ such that

$$\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1 \geq |w_i^* - \hat{w}_{s,i}| + |\hat{w}_{s,j}| \geq |w_i^*| \geq \min_{r\in S}|w_r^*|,$$

in which $|\hat{w}_{s,i}| \leq |\hat{w}_{s,j}|$. For any $s > s_2$, we have $2\frac{\mu_2 a_5 k}{\delta_S^2}\sqrt{\frac{d-1}{s(k-1)}} < \min_{i\in S}|w_i^*|$, which implies that, with probability at least $1 - \sqrt{T}\left(6 + \log_{1.5}\frac{d+k}{k-2}\right)\delta$,

$$\|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1 \leq 2\|\hat{\mathbf{w}}_s(S) - \mathbf{w}^*\|_1 < \min_{i\in S}|w_i^*| \leq \|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1.$$

There is a contradiction. We conclude the proof. ■

Proof [of Theorem 1] Let $s_T = \lfloor \sqrt{T} \rfloor$ and $T > (s_2+1)^2 > (s_1+1)^2$. For simplicity, we will alternately use the notation $\hat{y}_t = \langle \bar{\mathbf{w}}_t(S_s), \mathbf{x}_t \rangle$. The regret of DS-OSLRC can be decomposed as follows,

$$\begin{aligned}\mathrm{Reg}(\mathbf{w}^*) =& \underbrace{\sum_{s=1}^{s_T} [\ell(\hat{y}_{s^2}, y_{s^2}) - \ell(\langle \mathbf{w}^*, \mathbf{x}_{s^2} \rangle, y_{s^2})]}_{:=\mathrm{Reg}_0} + \sum_{s=1}^{s_T} \sum_{t=s^2+1}^{(s+1)^2-1} [\ell(\hat{y}_t, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t)] \\ =& \mathrm{Reg}_0 + \left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{s_2} + \sum_{s=s_2+1}^{s_T} \right) \sum_{t \in \mathcal{T}_s} \left[\ell(\langle \bar{\mathbf{w}}_t(S_s), \mathbf{x}_t \rangle, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t) \right],\end{aligned}$$

in which $(s_T+1)^2 - 1 := T$. For Reg_0, unfolding the square loss function and substituting into the definition of y_{s^2} gives

$$\sum_{s=1}^{s_T} \ell(\hat{y}_{s^2}, y_{s^2}) - \ell(\langle \mathbf{w}^*, \mathbf{x}_{s^2} \rangle, y_{s^2}) = \sum_{s=1}^{s_T} (\langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle)^2 - 2\eta_{s^2} \langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle.$$

Next we separately analyze the regret in the intervals $[2, (s_0+1)^2-1]$, $[(s_0+1)^2+1, (s_1+1)^2-1]$, $[(s_1+1)^2+1, (s_2+1)^2-1]$, and $[(s_2+1)^2+1, T]$.

We briefly explain the novelty of our analysis. As our algorithm updates $\bar{\mathbf{w}}_t(S_s)$ by ONS, the regret in each $\mathcal{T}_s$ can be decomposed into two components. The first one is the regret of our algorithm w.r.t. any $\mathbf{w}_s \in \mathcal{W}_s$. Note that we can arbitrary choose $\mathbf{w}_s \in \mathcal{W}_s$. Specifically, we will choose $\mathbf{w}_s$ by Lemma 5, i.e., $\mathbf{w}_s = \mathbf{w}^*(S_s \cap S)$. ONS ensures that our algorithm converges rapidly to $\mathbf{w}_s$. The second part of the regret is the difference of the cumulative losses of $\mathbf{w}_s$ and $\mathbf{w}^*$. As $\|\mathbf{w}_s - \mathbf{w}^*\|_1 \leq \|\hat{\mathbf{w}}_s - \mathbf{w}^*\|_1$, it is easy to analyze the second part by the convergence rate in Lemma 2. Compared with the vanilla selection of $\mathbf{w}_s$, i.e., $\mathbf{w}_s = \hat{\mathbf{w}}_s(S_s)$, our choice can significantly reduce the constant factor (please see the discussion below Lemma 5). With probability at least $1-\delta$, $\langle \mathbf{w}^*, \mathbf{x}_t \rangle \leq Y_\delta$ for a fixed t. Then with probability at least $1 - ((s_0+1)^2 - 1)\delta$,

$$\begin{aligned}&\mathrm{Reg}_{[1:s_0]}(\mathbf{w}^*) \\ =& \sum_{s=1}^{s_0} \sum_{t \in \mathcal{T}_s} [\ell(\hat{y}_t, y_t) - \ell(\langle \mathbf{w}_s, \mathbf{x}_t \rangle, y_t) + \ell(\langle \mathbf{w}_s, \mathbf{x}_t \rangle, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t)] \\ =& \sum_{s=1}^{s_0} \sum_{t \in \mathcal{T}_s} 2(\hat{y}_t - y_t)(\hat{y}_t - \langle \mathbf{w}_s, \mathbf{x}_t \rangle) - (\hat{y}_t - \langle \mathbf{w}_s, \mathbf{x}_t \rangle)^2 + (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \\ =& \sum_{s=1}^{s_0} \sum_{t \in \mathcal{T}_s} \langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle - \frac{(\langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle)^2}{4(\hat{y}_t - y_t)^2} + (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \\ \leq& \sum_{s=1}^{s_0} \sum_{t \in \mathcal{T}_s} \langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle - \frac{(\langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle)^2}{4(1+Y_\delta)^2} + (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle.\end{aligned}$$

By the regret analysis of ONS (Hazan et al., 2007), we have

$$\begin{aligned}\|\bar{\mathbf{w}}_{t+1}(S_s) - \mathbf{w}_s\|_{\mathbf{A}_t}^2 \leq& \left\|\bar{\mathbf{w}}_t(S_s) - \mathbf{A}_t^{-1}\mathbf{g}_t - \mathbf{w}_s\right\|_{\mathbf{A}_t}^2 \\ =& \|\bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s\|_{\mathbf{A}_t}^2 - 2\left\langle \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s, \mathbf{A}_t^{-1}\mathbf{g}_t \right\rangle_{\mathbf{A}_t} + \left\|\mathbf{A}_t^{-1}\mathbf{g}_t\right\|_{\mathbf{A}_t}^2.\end{aligned}$$

Recalling that $\rho = \frac{1}{2(1+Y_\delta)^2}$. Rearranging terms yields

$$
\begin{aligned}
&\sum_{t\in\mathcal{T}_s} \langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle - \frac{(\langle \mathbf{g}_t, \bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s \rangle)^2}{4(1+Y_\delta)^2} \\
\le &\sum_{t=s^2+1}^{(s+1)^2-1} \frac{\|\bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s\|_{\mathbf{A}_t}^2 - \|\bar{\mathbf{w}}_{t+1}(S_s) - \mathbf{w}_s\|_{\mathbf{A}_t}^2}{2} + \frac{\|\mathbf{A}_t^{-1}\mathbf{g}_t\|_{\mathbf{A}_t}^2}{2} - \frac{\|\bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s\|_{\rho\cdot\mathbf{g}_t\mathbf{g}_t^\top}^2}{2} \\
= &\sum_{t=s^2+1}^{(s+1)^2-1} \frac{\|\bar{\mathbf{w}}_t(S_s) - \mathbf{w}_s\|_{\mathbf{A}_{t-1}}^2 - \|\bar{\mathbf{w}}_{t+1}(S_s) - \mathbf{w}_s\|_{\mathbf{A}_t}^2}{2} + \frac{\|\mathbf{A}_t^{-1}\mathbf{g}_t\|_{\mathbf{A}_t}^2}{2} \\
= &\underbrace{\frac{\|\bar{\mathbf{w}}_{s^2+1}(S_s) - \mathbf{w}_s\|_{\mathbf{A}_{s^2}}^2}{2} - \frac{\|\bar{\mathbf{w}}_{(s+1)^2}(S_s) - \mathbf{w}_s\|_{\mathbf{A}_{(s+1)^2-1}}^2}{2} + \sum_{t=s^2+1}^{(s+1)^2-1} \frac{\mathbf{g}_t^\top \mathbf{A}_t^{-1}\mathbf{g}_t}{2}}_{:=\Xi_s},
\end{aligned}
$$

where $\mathbf{A}_t = \mathbf{A}_{t-1} + \rho \cdot \mathbf{g}_t\mathbf{g}_t^\top$. We have

$$
\mathrm{Reg}_{[1:s_0]}(\mathbf{w}^*) \le \sum_{s=1}^{s_0} \Xi_s + \sum_{s=1}^{s_0} \left[\sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \right].
$$

Similarly, the regret in the intervals $[(s_0+1)^2+1, (s_1+1)^2-1]$, $[(s_1+1)^2+1, (s_2+1)^2-1]$, and $[(s_2+1)^2+1, T]$ can be decomposed as follows,

$$
\begin{aligned}
\mathrm{Reg}_{[s_0:s_1]}(\mathbf{w}^*) &:= \sum_{s=s_0+1}^{s_1} \sum_{t=s^2+1}^{(s+1)^2-1} [\ell(\hat{y}_t, y_t) - \ell(\langle \mathbf{w}_s, \mathbf{x}_t \rangle, y_t) + \ell(\langle \mathbf{w}_s, \mathbf{x}_t \rangle, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t)] \\
&= \sum_{s=s_0+1}^{s_1} \Xi_s + \sum_{s=s_0+1}^{s_1} \left[\sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \right], \\
\mathrm{Reg}_{[s_1:s_2]}(\mathbf{w}^*) &:= \sum_{s=s_1+1}^{s_2} \sum_{t=s^2+1}^{(s+1)^2-1} [\ell(\hat{y}_t, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t)] \\
&\le \sum_{s=s_1+1}^{s_2} \Xi_s + \sum_{s=s_1+1}^{s_2} \left[\sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \right], \\
\mathrm{Reg}_{[s_2:s_T]}(\mathbf{w}^*) &:= \sum_{s=s_2+1}^{s_T} \sum_{t=s^2+1}^{(s+1)^2-1} [\ell(\hat{y}_t, y_t) - \ell(\langle \mathbf{w}^*, \mathbf{x}_t \rangle, y_t)] \\
&\le \sum_{s=s_2+1}^{s_T} \Xi_s + \sum_{s=s_2+1}^{s_T} \left[\sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 - 2\eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle \right].
\end{aligned}
$$

Summing over all results gives, with probability at least $1 - T\delta$,

$$\mathrm{Reg}(\mathbf{w}^*) =$$
$$\underbrace{\left[\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{s_2} + \sum_{s=s_2+1}^{s_T}\right] \frac{\|\bar{\mathbf{w}}_{s^2+1}(S_s) - \mathbf{w}_s\|^2_{\mathbf{A}_{s^2}} - \|\bar{\mathbf{w}}_{(s+1)^2}(S_s) - \mathbf{w}_s\|^2_{\mathbf{A}_{(s+1)^2-1}}}{2}}_{:=\mathrm{Reg}_1} +$$
$$\underbrace{\left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{s_2} + \sum_{s=s_2+1}^{s_T}\right) \sum_{t=s^2+1}^{(s+1)^2-1} \frac{\mathbf{g}_t^\top \mathbf{A}_t^{-1} \mathbf{g}_t}{2}}_{:=\mathrm{Reg}_2} +$$
$$\underbrace{\left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{s_2} + \sum_{s=s_2+1}^{s_T}\right) \sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 + \sum_{s=1}^{s_T} (\langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle)^2}_{:=\mathrm{Reg}_3} +$$
$$\underbrace{-2\left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{s_2} + \sum_{s=s_2+1}^{s_T}\right) \sum_{t=s^2+1}^{(s+1)^2-1} \eta_t \langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle - \sum_{s=1}^{s_T} 2\eta_{s^2} \langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle}_{:=\mathrm{Reg}_4}.$$

We first analyze Reg_4. By the concentration inequality of Gaussian variables, with probability at least $1 - \delta$,

$$\mathrm{Reg}_4 \leq 2\sqrt{2\mathrm{Reg}_3 \cdot \ln \frac{1}{\delta}}.$$

Next we analyze Reg_3.

$$\sum_{s=1}^{s_T} (\langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle)^2 \leq s_T \cdot \|\hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*\|_1^2 \cdot \|\mathbf{x}_{s^2}\|_\infty^2 \leq 4s_T,$$
$$\sum_{s=1}^{s_0} \sum_{t=s^2+1}^{(s+1)^2-1} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 \leq 4s_0(s_0+1).$$

Next we analyze the regret in the intervals $[(s_0+1)^2+1, (s_1+1)^2-1]$, $[s_1^2+1, (s_2+1)^2-1]$ and $[(s_2+1)^2+1, T]$. For any $s \geq 2$, if $S_s \neq S_{s-1}$, then we call such a round "breakpoint". Assuming that there are n breakpoints in the set $\{2, 3, \ldots, s_T\}$, denoted by $b_1, b_2, \ldots, b_n$. Specifically, for each $j = 1, 2, \ldots, n$, we have $S_{b_j} \neq S_{b_j - 1}$. Without loss of generality, assuming that there are two breakpoints $b_r, b_m \in \{b_1, b_2, \ldots, b_n\}$ such that

$$b_{r-1} < s_0 + 1 < b_r, \quad b_{m-1} < s_1 + 1 < b_m.$$

By Lemma 17, under the condition that $\Delta_s(S)$ satisfies Lemma 2 for any $s \geq 1$, it must be

$$\forall s \geq s_2 + 1, \quad S_s = S_{s+1} = \ldots = S_{s_T} = S,$$

which means there is no breakpoints in the interval $[s_2+2, s_T]$ and

$$b_n \le s_2+1, \quad n \le s_2. \tag{21}$$

Next we define the competitor $\mathbf{w}_s$ as follows.

$$\begin{cases} \forall s \in \{s_0+1, \dots, b_r-1\}, & \mathbf{w}_s = \mathbf{w}^*(S_{b_r-1} \cap S) := \mathbf{w}^*_{b_r-1}, \\ \forall s \in \{b_\tau, b_{\tau+1}-1\}, r \le \tau \le n-2, & \mathbf{w}_s = \mathbf{w}^*(S_{b_{\tau+1}-1} \cap S) := \mathbf{w}^*_{b_{\tau+1}-1}, \\ \forall s \in \{b_{n-1}, \dots, s_2\}, & \mathbf{w}_s = \mathbf{w}^*(S_{s_2} \cap S) := \mathbf{w}^*_{s_2}, \\ \forall s \in \{b_n, \dots, s_T\}, & \mathbf{w}_s = \mathbf{w}^*. \end{cases} \tag{22}$$

By Lemma 5, (13) and (21),

$$\begin{aligned}
&\mathrm{Reg}_3 - \sum_{s=1}^{s_T} (\langle \hat{\mathbf{w}}_{s-1}(B_s) - \mathbf{w}^*, \mathbf{x}_{s^2} \rangle)^2 - \sum_{s=1}^{s_0} \sum_{t \in \mathcal{T}_s} (\langle \mathbf{w}_s - \mathbf{w}^*, \mathbf{x}_t \rangle)^2 \\
\le& \left(\sum_{s=s_0+1}^{b_r-1} + \sum_{\tau=r}^{n-1} \sum_{s=b_\tau}^{b_{\tau+1}-1} + \sum_{s=b_n}^{s_T} \right) \sum_{t=s^2+1}^{(s+1)^2-1} \|\mathbf{w}_s - \mathbf{w}^*\|_1^2 \cdot \|\mathbf{x}_t\|_\infty^2 \\
\le& \left(\sum_{s=s_0+1}^{b_r-1} + \sum_{\tau=r}^{n-2} \sum_{s=b_\tau}^{b_{\tau+1}-1} + \sum_{s=b_{n-1}}^{s_2} \right) \sum_{t=s^2+1}^{(s+1)^2-1} \|\mathbf{w}_s - \mathbf{w}^*\|_1^2 \cdot \|\mathbf{x}_t\|_\infty^2 \\
\le& 2 \left(\sum_{s=s_0+1}^{b_r-1} \|\mathbf{w}^*_{b_r-1} - \mathbf{w}^*\|_1^2 + \sum_{\tau=r}^{n-2} \sum_{s=b_\tau}^{b_{\tau+1}-1} \|\mathbf{w}^*_{b_{\tau+1}-1} - \mathbf{w}^*\|_1^2 + \sum_{s=b_{n-1}}^{s_2} \|\mathbf{w}^*_{s_2} - \mathbf{w}^*\|_1^2 \right) \cdot s \\
\le& 2 \left(\sum_{s=s_0+1}^{b_r-1} \|\hat{\mathbf{w}}_{b_r-1} - \mathbf{w}^*\|_1^2 + \sum_{\tau=r}^{n-2} \sum_{s=b_\tau}^{b_{\tau+1}-1} \|\hat{\mathbf{w}}_{b_{\tau+1}-1} - \mathbf{w}^*\|_1^2 + \sum_{s=b_{n-1}}^{s_2} \|\hat{\mathbf{w}}_{s_2} - \mathbf{w}^*\|_1^2 \right) \cdot s \\
\le& 8 \left(\sum_{s=s_0+1}^{b_r-1} \|\Delta_{b_r-1}(S)\|_1^2 + \sum_{\tau=r}^{n-2} \sum_{s=b_\tau}^{b_{\tau+1}-1} \|\Delta_{b_{\tau+1}-1}(S)\|_1^2 + \sum_{s=b_{n-1}}^{s_2} \|\Delta_{s_2}(S)\|_1^2 \right) \cdot s \\
\le& 8 \sum_{s=s_0+1}^{s_1} \|\Delta_s(S)\|_1^2 \cdot s + 8 \sum_{s=s_1+1}^{s_2} \|\Delta_s(S)\|_1^2 \cdot s \qquad \text{(by Lemma 2)} \\
\le& 22 \frac{a_4^2}{\delta_S^8} k^4 g_{d,k}^2 \ln \frac{s_1}{s_0} + 8 \frac{(\mu_2 a_5)^2}{\delta_S^4} \frac{k^2(d-1)}{k-1} (s_2 - s_1).
\end{aligned}$$

Summing all results gives

$$\begin{aligned}
\mathrm{Reg}_3 \le& 4 s_T + 4 s_0 (s_0+1) + 22 \frac{a_4^2}{\delta_S^8} k^4 g_{d,k}^2 \ln \frac{s_1}{s_0} + 8 \cdot \frac{(\mu_2 a_5)^2}{\delta_S^4} \frac{k^2(d-1)}{k-1} (s_2 - s_1) \\
\le& 4\sqrt{T} + 22 \frac{a_4^2}{\delta_S^8} k^4 g_{d,k}^2 \ln \frac{(d-2) \ln \frac{d}{\delta}}{k-2} + 2 \frac{(2\mu_2 a_5)^4}{\delta_S^8} \cdot \frac{k^4 (d-1)^2}{\min_{i \in S} |w_i^*|^2 (k-1)^2}.
\end{aligned}$$

Now we begin to analyze Reg_1. For $s \in [1, b_1 - 1]$, we define

$$\mathbf{w}_s = \mathbf{w}^*(S_{b_1-1} \cap S) := \mathbf{w}^*_{b_1-1}.$$

Recalling (22), for $\tau \in [b_\tau, b_{\tau+1}-1]$, in which $1 \le \tau \le n-1$, we define $\mathbf{w}_s = \mathbf{w}^*_{b_{\tau+1}-1}$. For $s \in [b_n, s_T]$, $\mathbf{w}_s = \mathbf{w}^*$.

$$
\begin{aligned}
\mathrm{Reg}_1 =& \sum_{s=1}^{b_1-1}\left[\frac{\|\bar{\mathbf{w}}_{s^2+1}(S_s)-\mathbf{w}^*_{b_1-1}\|^2_{\mathbf{A}_{s^2}}}{2}-\frac{\|\bar{\mathbf{w}}_{(s+1)^2}(S_s)-\mathbf{w}^*_{b_1-1}\|^2_{\mathbf{A}_{(s+1)^2-1}}}{2}\right]+\\
&\sum_{r=1}^{n-1}\sum_{s=b_r}^{b_{r+1}-1}\frac{\|\bar{\mathbf{w}}_{s^2+1}(S_s)-\mathbf{w}^*_{b_{r+1}-1}\|^2_{\mathbf{A}_{s^2}}}{2}-\frac{\|\bar{\mathbf{w}}_{(s+1)^2}(S_s)-\mathbf{w}^*_{b_{r+1}-1}\|^2_{\mathbf{A}_{(s+1)^2-1}}}{2}+\\
&\sum_{s=b_n}^{s_T}\left[\frac{\|\bar{\mathbf{w}}_{s^2+1}(S_s)-\mathbf{w}^*\|^2_{\mathbf{A}_{s^2}}}{2}-\frac{\|\bar{\mathbf{w}}_{(s+1)^2}(S_s)-\mathbf{w}^*\|^2_{\mathbf{A}_{(s+1)^2-1}}}{2}\right]+\\
\le&\frac{\|\bar{\mathbf{w}}_2(S_1)-\hat{\mathbf{w}}^*_{b_1-1}\|^2_{\mathbf{A}_1}}{2}+\sum_{r=1}^{n-1}\frac{\|\bar{\mathbf{w}}_{(b_r)^2+1}(S_{b_r})-\hat{\mathbf{w}}^*_{b_{r+1}-1}\|^2_{\mathbf{A}_{(b_r)^2}}}{2}+\\
&\frac{\|\bar{\mathbf{w}}_{(b_n)^2+1}(S_s)-\mathbf{w}^*\|^2_{\mathbf{A}_{(b_n)^2}}}{2}\\
\le&4\varepsilon+2(n-1)\varepsilon,
\end{aligned}
$$

in which $\bar{\mathbf{w}}_{s^2+1}(S_s) = \bar{\mathbf{w}}_{s^2}(S_{s-1})$ and $\mathbf{A}_{s^2} = \mathbf{A}_{s^2-1}$ for all s such that $S_s = S_{s-1}$, and $\mathbf{A}_{s^2} = \varepsilon \cdot \mathbf{I}_{k\times k}$ for $s = 1, b_1, b_2, \ldots, b_n$.

Finally, we analyze Reg_2. By Lemma 8,

$$
\begin{aligned}
\mathrm{Reg}_2 =& \sum_{s=1}^{b_1-1}\sum_{t=s^2+1}^{(s+1)^2-1}\frac{\mathbf{g}_t^\top\mathbf{A}_t^{-1}\mathbf{g}_t}{2}+\sum_{r=1}^{n-1}\sum_{s=b_r}^{b_{r+1}-1}\sum_{t=s^2+1}^{(s+1)^2-1}\frac{\mathbf{g}_t^\top\mathbf{A}_t^{-1}\mathbf{g}_t}{2}+\sum_{s=b_n}^{s_T}\sum_{t=s^2+1}^{(s+1)^2-1}\frac{\mathbf{g}_t^\top\mathbf{A}_t^{-1}\mathbf{g}_t}{2}\\
\le&\frac{nk}{2}\ln\left(\frac{4(1+Y_\delta)^2kb_n^2}{\varepsilon}+1\right)+\frac{k}{2}\ln\left(\frac{4(1+Y_\delta)^2kT}{\varepsilon}+1\right),
\end{aligned}
$$

in which $\|\mathbf{g}_t\|_2 \le 2\left(1+Y_\delta\right)\|\mathbf{x}_t(S_s)\|_2 \le 2\left(1+Y_\delta\right)\sqrt{k}$. By (21),

$$
\mathrm{Reg}_1+\mathrm{Reg}_2 \le 2s_2\varepsilon+\frac{s_2}{2}k\ln\left(\frac{4(1+Y_\delta)^2k(s_2+1)^2}{\varepsilon}+1\right)+\frac{k}{2}\ln\left(\frac{4(1+Y_\delta)^2kT}{\varepsilon}+1\right).
$$

Let $\varepsilon = k$. Summing all results gives, with probability at least $1-(T+\sqrt{T}(6+\log_{1.5}\frac{2(d-2)}{3(k-2)})+1)\delta$, we have

$$
\begin{aligned}
\mathrm{Reg}(\mathbf{w}^*) \le& 4\sqrt{T}+2ks_2+\frac{s_2k}{2}\ln\left(4(1+Y_\delta)^2s_2^2+1\right)+\frac{k}{2}\ln\left(4(1+Y_\delta)^2T+1\right)\\
&\frac{22a_4^2}{\delta_S^8}k^4g_{d,k}^2\ln\frac{(d-2)\ln\frac{d}{\delta}}{k-2}+\frac{2(2\mu_2a_5)^4}{\delta_S^8}\frac{k^4(d-1)^2}{\min_{i\in S}|w_i^*|^2(k-1)^2}+2\sqrt{2\mathrm{Reg}_3\ln\frac{1}{\delta}}.
\end{aligned}
$$

Omitting the lower order terms concludes the proof. ■

Appendix E. Proof of Theorem 2

We first prove a technical lemma similar to Lemma 5.

Lemma 18 *For any $s \geq 1$, let $\hat{\mathbf{w}}_s$ be the solution of $\mathrm{DS}(\hat{\gamma}_s)$. Let $S_s \subseteq [d]$ satisfy $|S_s| = k$ and for any $i \in S_s$ and $j \in [d] \setminus S_s$, $|\hat{w}_{s,i}| \geq |\hat{w}_{s,j}|$. If $\hat{\gamma}_\tau \geq \gamma_\tau$ for all $\tau \leq s$, then w.p. at least $1 - s\left(5 + \log_{1.5} \frac{2(d'-2)}{3(k_0-2)}\right)\delta$,*

$$\forall \tau \leq s, \quad \|\hat{\mathbf{w}}_\tau(S_\tau) - \mathbf{w}^*\|_1 \leq 3\|\hat{\mathbf{w}}_\tau(S) - \mathbf{w}^*\|_1.$$

Proof [of Lemma 18] Unfolding $\|\mathbf{w}^*_\tau(S_\tau) - \mathbf{w}^*\|_1$ gives

$$\begin{aligned}
\|\hat{\mathbf{w}}_\tau(S_\tau) - \mathbf{w}^*\|_1 =& \sum_{i \in S \cap S_\tau} |\hat{w}_{\tau,i} - w^*_i| + \sum_{i \in S_\tau \setminus S} |\hat{w}_{\tau,i}| + \sum_{i \in S \setminus S_\tau} |w^*_i| \\
=& \sum_{i \in S \cap S_\tau} |\hat{w}_{\tau,i} - w^*_i| + \sum_{i \in S_\tau \setminus S} |\hat{w}_{\tau,i}| + \sum_{i \in S \setminus S_\tau} |w^*_i - \hat{w}_{\tau,i} + \hat{w}_{\tau,i}| \\
\leq& \sum_{i \in S \cap S_\tau} |\hat{w}_{\tau,i} - w^*_i| + \sum_{i \in S_\tau \setminus S} |\hat{w}_{\tau,i}| + \sum_{i \in S \setminus S_\tau} |w^*_i - \hat{w}_{\tau,i}| + \sum_{i \in S \setminus S_\tau} |\hat{w}_{\tau,i}| \\
\leq& \|\hat{\mathbf{w}}_\tau(S) - \mathbf{w}^*\|_1 + 2 \sum_{i \in S_\tau \setminus S} |\hat{w}_{\tau,i}| \\
\leq& 3\|\hat{\mathbf{w}}_\tau(S) - \mathbf{w}^*\|_1,
\end{aligned}$$

where the last inequality comes from Lemma 16. We conclude the proof. ■

Proof [of Theorem 2] For simplicity, we will alternately use the notation $\hat{y}_s = \langle \hat{\mathbf{w}}_{s-1}(S_{s-1}), \mathbf{x}_s(S_{s-1}) \rangle$. With probability at least $1 - (T(6 + \log_{1.5} \frac{d+k}{k-2}) + 1)\delta$, the regret of DS-POSLRC satisfies

$$\begin{aligned}
&\mathrm{Reg}(\mathbf{w}^*) \\
=& \left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{T} \right) \left[\ell\left(\langle \hat{\mathbf{w}}_{s-1}(S_{s-1}), \mathbf{x}_s(S_{s-1}) \rangle, y_s \right) - \ell\left(\langle \mathbf{w}^*, \mathbf{x}_s \rangle, y_s \right) \right] \\
=& \left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{T} \right) \left[(\langle \hat{\mathbf{w}}_{s-1}(S_{s-1}) - \mathbf{w}^*, \mathbf{x}_s \rangle)^2 - 2\eta_t \langle \hat{\mathbf{w}}_{s-1}(S_{s-1}) - \mathbf{w}^*, \mathbf{x}_s \rangle \right] \\
\leq& \left(\sum_{s=1}^{s_0} + \sum_{s=s_0+1}^{s_1} + \sum_{s=s_1+1}^{T} \right) \|\hat{\mathbf{w}}_{s-1}(S_{s-1}) - \mathbf{w}^*\|_1^2 + 2\sqrt{2 \sum_{s=1}^{T} \|\hat{\mathbf{w}}_{s-1}(S_{s-1}) - \mathbf{w}^*\|_1^2 \cdot \ln \frac{1}{\delta}} \\
\leq& \frac{48^2 k^2 g'_{d',k_0}}{\delta_S^4} \ln \frac{d^2}{\delta} + \frac{2.7 a_4^2 k^2 g_{d',k_0}}{64 \delta_S^4 \ln \frac{d^2}{\delta}} + 9\mu_2^2 a_5^2 \frac{k^2(d'-1)}{\delta_S^4 (k_0 - 1)} \ln \frac{T}{s_1 + 1} + \\
& 2\sqrt{2 \left(\frac{48^2 k^2 g'_{d',k_0}}{\delta_S^4} \ln \frac{d^2}{\delta} + \frac{2.7 a_4^2 k^2 g_{d',k_0}}{64 \delta_S^4 \ln \frac{d^2}{\delta}} + 9\mu_2^2 a_5^2 \frac{k^2(d'-1)}{\delta_S^4 (k_0 - 1)} \ln \frac{T}{s_1 + 1} \right) \ln \frac{1}{\delta}},
\end{aligned}$$

in which by Lemma 3 and Lemma 18, we have

$$\begin{aligned}
&\left(\sum_{s=1}^{s_0}+\sum_{s=s_0+1}^{s_1}+\sum_{s=s_1+1}^{T}\right)\|\hat{\mathbf{w}}_{s-1}(S_{s-1})-\mathbf{w}^*\|_1^2\\
\leq& 4s_0+9\left(\sum_{s=s_0+1}^{s_1}+\sum_{s=s_1+1}^{T}\right)\|\hat{\mathbf{w}}_{s-1}(S)-\mathbf{w}^*\|_1^2\\
\leq& 4s_0+9\left(\frac{9}{9-2\sqrt{3}}\right)^2\frac{a_4^2k^4g_{d',k_0}^2}{\delta_S^8}\left(\frac{1}{s_0}-\frac{1}{s_1}\right)+9\mu_2^2a_5^2\frac{k^2(d'-1)}{\delta_S^4(k_0-1)}\ln\frac{T}{s_1+1}\\
\leq& 4s_0+\frac{2.7a_4^2k^2g_{d',k_0}}{64\delta_S^4\ln\frac{d^2}{\delta}}+9\mu_2^2a_5^2\frac{k^2(d'-1)}{\delta_S^4(k_0-1)}\ln\frac{T}{s_1+1}.
\end{aligned}$$

Omitting the lower order terms concludes the proof. ■