# Regression Learning with Limited Observations of Multivariate Outcomes and Features

**Yifan Sun** [1] **Grace Y. Y** [1 2]

## Abstract

Multivariate linear regression models are broadly used to facilitate relationships between outcomes and features. However, their effectiveness is compromised by the presence of missing observations, a ubiquitous challenge in real-world applications. Considering a scenario where learners access only limited components for both outcomes and features, we develop efficient algorithms tailored for the least squares ($L_2$) and least absolute ($L_1$) loss functions, each coupled with a ridge-like and Lasso-type penalty, respectively. Moreover, we establish rigorous error bounds for all proposed algorithms. Notably, our $L_2$ loss function algorithms are probably approximately correct (PAC), distinguishing them from their $L_1$ counterparts. Extensive numerical experiments show that our approach outperforms methods that apply existing algorithms for univariate outcome individually to each coordinate of multivariate outcomes in a naive manner. Further, utilizing the $L_1$ loss function or introducing a Lasso-type penalty can enhance predictions in the presence of outliers or high dimensional features. This research contributes valuable insights into addressing the challenges posed by incomplete data.

## 1. Introduction

Datasets with multiple outcomes are pervasive in many practical applications, especially in the era of big data. Typical examples include longitudinal data (Diggle, 2002), panel data (Baltagi & Baltagi, 2008), functional data (Horváth & Kokoszka, 2012), and image data (Zhu et al., 2023). Multivariate linear regression is a simple and widely used tool to characterize relationships between multi-dimensional outcomes and multiple features. It is commonly employed to predict outcomes for new features. The literature on this topic is extensive; some notable works include Breiman & Friedman (1997); Su et al. (2012); Price & Sherwood (2018), and the references therein.

However, the applicability of those available methods is often hindered by limitations in observations, a phenomenon arising frequently in applications. This constraint can occur intentionally as part of a design. For example, in machine learning tasks involving high-dimensional data like electronic health records and large-scale administrative data, selecting a subset of features and/or outcomes is crucial due to constraints such as limited computational resources.Moreover, not all variables contribute equally to predictive performance. Opting for the most informative variables not only enhances the interpretability of the model but also improves the model generalization ability by mitigating the risk of overfitting.

On the other hand, limited observations can arise beyond our control. For example, in medical studies, not all examination items for each patient can be measured due to factors like time constraints or cost. Similarly, in weather forecasting, it is difficult or even impossible to observe the entire historical data, necessitating the selection of representative stations and specific times for measurements. These constraints drive the need for algorithms capable of handling datasets with restricted observations.

Under the univariate outcome setting, the issue of missing features has been extensively studied in statistics and received attention in the machine learning literature. This problem is also named as learning with *limited attribute observation* (LAO) (Ben-David & Dichterman, 1993), *restricted focus of attention* (Ben-David & Dichterman, 1998), or *budget learning* (Madani et al., 2004). Cesa-Bianchi et al. (2011) analyzed the linear prediction in this setting and proposed algorithms with theoretical guarantees for the generalization error. Hazan & Koren (2012) proposed more efficient online algorithms for regression with ridge or Lasso constraint, which was then improved by Kukliansky & Shamir (2015) using a different sampling strategy. Rostamizadeh et al. (2011) developed batch and online algo-

[1]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Canada [2]Department of Computer Science, University of Western Ontario, London, Canada. Correspondence to: Grace Y. Y <gyi5@uwo.ca>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

rithms to simultaneously learn the imputation and prediction functions. The computational efficiency for sequential prediction using a limited number of features was explored by Foster et al. (2016) and Kale et al. (2017). Classification with missing features was studied by Dekel & Shamir (2008); Hazan et al. (2015); Gong et al. (2023), among others.

However, when dealing with multi-dimensional outcomes, much of the existing research has predominantly focused on addressing missing outcomes alone (e.g., Ibrahim & Molenberghs (2009), and the reference therein). Surprisingly, there has been relatively limited exploration into scenarios where both outcomes and features are missing simultaneously, especially in settings with multiple outcomes. In practice, however, data are often unavailable for both outcomes and features.

Chen et al. (2008) explored this scenario using a likelihood-based method and investigated the theoretical properties within the normal linear model. Chen et al. (2010) studied intermittently missing-at-random data and proposed an estimation equation approach. Addressing multi-view multi-label data, Tan et al. (2018) developed algorithms to learn from both incomplete views and weak labels. For multi-label learning with incomplete binary labels and features, Han et al. (2018) and Hao et al. (2022) proposed methods to simultaneously recover missing variables and learn classification models.

### 1.1. Related Works and Our Contributions

Previous research on LAO has primarily focused on scenarios with univariate outcomes but has not explored other settings. To address this significant gap, our article investigates a broader context where the outcome is multi-dimensional, allowing for LAO to occur in both features and outcome variables. Our study centers on a unique missingness scenario where learners can select a limited set of components from both outcomes and features for observation. This framework encompasses the settings explored by Cesa-Bianchi et al. (2011),Hazan & Koren (2012), and Bullins et al. (2016), who focused on the univariate outcome setting. Our work uncovers new findings that generalize the results from those studies.

Table 1 presents a comparison of those related works, with $T$, $p$, and $q$ denoting the sample size, feature dimension, and outcome dimension, respectively. Here $p_0$ and $q_0$ represent the number of observed features and observed outcomes, respectively. Under the 'LAO' column, 'y,x' and 'x' represent scenarios with missing attributes in both outcomes and features, and missing features only, respectively. The last two columns of the table display the (squared) Excess Error Bound (EEB) under least squares and least absolute loss functions. In the table, '$O$' denotes the expected EEB, while '$O_\delta$' represents the EEB with a probability of $1-\delta$, up to a constant factor.

*Table 1.* A summary of our work as opposed to related methods concerning LAO. Here, C, H, and B represent Cesa-Bianchi et al. (2011), Hazan & Koren (2012), and Bullins et al. (2016), respectively.

| WORK | $q$ | LAO | EEB ($L_2$) | EEB ($L_1$) |
|---|---|---|---|---|
| OURS | $\geq 1$ | $y, x$ | $O\left(\frac{pq}{T(p_0-1)q_0}\right)$ | $O\left(\frac{q^2 p_0}{pq_0 T}+\frac{q(p-q_0)}{p}\right)$ |
| C | 1 | $x$ | $O_\delta\left(\frac{p}{p_0 T}\log\frac{T}{\delta}\right)$ | - |
| H | 1 | $x$ | $O\left(\frac{p}{T(p_0-1)}\right)$ | - |
| B | 1 | $x$ | - | $O\left(\frac{1}{T}+\frac{p-1}{p}\right)$ |

Our research represents a notable departure from existing literature in the following key aspects:

- Our work broadens the scope of existing research in this area by generalizing the univariate outcome framework to include multi-dimensional outcomes. Further, our work accommodates LAO occurring in both features and outcomes.
- In contrast to much of the statistical literature, which typically emphasizes statistical inference and often assumes a missing-at-random mechanism (Chen et al., 2008; 2010), our work stands out for its emphasis on prediction properties. We do not rely on specific regression models or missing data models before employing the developed algorithms. This aspect makes our approach particularly attractive, given the challenges associated with validating such models in practical applications.
- We extend beyond the least squares loss function to incorporate the least absolute loss function to enhance prediction robustness. We introduce efficient algorithms for different loss functions under ridge-like constraints or an additional Lasso-type penalty. Furthermore, we rigorously establish expected risk bounds for the outputs of the proposed algorithms. Notably, while arbitrary accuracy can be achieved with a sufficiently large sample size in regression with a least squares loss function, a gap exists in regression with a least absolute loss function unless the features are fully observed. As shown in Table 1, our results align with univariate counterparts (Hazan & Koren, 2012; Bullins et al., 2016), with a slight improvement factor $p_0/p$ in the least absolute loss function setting.

In summary, our paper offers valuable new perspectives and insights into addressing Limited Attribute Observation,

which represents a distinct category within missing data problems.

## 2. Preliminaries

We use column vectors $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^q$ to denote features and outcomes that have dimensions $p$ and $q$, respectively. In multivariate linear regression, the learner aims at seeking a $p \times q$ weight matrix $\boldsymbol{W}$ such that the linear predictor $\hat{\boldsymbol{y}} \triangleq \boldsymbol{W}^\mathsf{T}\boldsymbol{x}$ provides a good prediction of $\boldsymbol{y}$. Let $L(\hat{\boldsymbol{y}}, \boldsymbol{y})$ denote a loss function, mapping from $\mathbb{R}^q \times \mathbb{R}^q$ to $\mathbb{R}$, which is convex in the first argument. Commonly used loss functions include half of the $L_2$ loss: $L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{2}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|^2$, and the $L_1$ loss: $L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_1$, where $\|\boldsymbol{a}\|$ and $\|\boldsymbol{a}\|_1$ represent the $L_2$ norm and the $L_1$ norm for vector $\boldsymbol{a}$, respectively. It is well-known that the least absolute deviation method, which utilizes the $L_1$ loss, produces estimates that are more robust to outliers than those obtained using the $L_2$ loss (Watt et al., 2020, Section 5.3)

Define the risk of the weight $\boldsymbol{W}$

$$R(\boldsymbol{W}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}}\{L(\boldsymbol{W}^\mathsf{T}\boldsymbol{x}, \boldsymbol{y})\}, \tag{1}$$

where $\mathcal{D}$ is the unknown distribution of $(\boldsymbol{x}, \boldsymbol{y})$ over $\mathbb{R}^p \times \mathbb{R}^q$. To learn a linear regressor, we often regularize $\boldsymbol{W}$ to minimize $R(\boldsymbol{W})$, with certain constraints on $\boldsymbol{W}$ imposed. Taking the regularization term to be the Frobenius norm of $\boldsymbol{W}$ yields the ridge-like regression

$$\operatorname*{argmin}_{\|\boldsymbol{W}\|_F \le B} R(\boldsymbol{W}), \tag{2}$$

where $\|\boldsymbol{A}\|_F$ represents the Frobenius norm for a matrix $\boldsymbol{A}$, and $B > 0$ is the regularization parameter (Hoerl & Kennard, 1970).

To address various data features, it is common to further introduce a penalty function for $\boldsymbol{W}$ to balance its trade off with $R(\boldsymbol{W})$. In other words, (2) can be extended to include a penalty function $P_\lambda(\boldsymbol{W})$ with tuning parameter $\lambda$:

$$\operatorname*{argmin}_{\|\boldsymbol{W}\|_F \le B} \{R(\boldsymbol{W}) + P_\lambda(\boldsymbol{W})\}. \tag{3}$$

Typically, $P_\lambda(\boldsymbol{W})$ is not chosen to be the Frobenius norm to avoid imposing redundant constriaints. Setting $P_\lambda(\boldsymbol{W})$ to be the $L_1$ norm of the vectorized $\boldsymbol{W}$ yields the lasso penalty (Tibshirani, 1996), which has many variants, including group Lasso (Yuan & Lin, 2006; Jacob et al., 2009) and sparse group Lasso (Simon et al., 2013; Ida et al., 2019), among others. We opt to keep this constraint as in (2) for the sake of convenience in the technical proof.

## 3. Algorithms for Least Squares

Let $\{(\boldsymbol{x}_t, \boldsymbol{y}_t) : t = 1, \ldots, T\}$ be independently and identically distributed samples of $(\boldsymbol{x}, \boldsymbol{y})$, where $T$ is the number of examples. Suppose that there exist positive constants $B_x$ and $B_y$ such that

$$\|\boldsymbol{x}_t\| \le B_x \text{ and } \|\boldsymbol{y}_t\| \le B_y \text{ for all } t. \tag{4}$$

We consider the setting with online data, for which we have access to $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ only at each time $t$. The learning algorithm is designed to predict $\boldsymbol{y}_t$ based on $\boldsymbol{x}_t$ and the previous information $\{(\boldsymbol{x}_j, \boldsymbol{y}_j) : j = 1, \ldots, t-1\}$.

We consider the case where both outcomes and features can only be limitedly observed. To be specific, although we may have access to full data, we are required to choose no more than $q_0$ coordinates of $\boldsymbol{y}$ and $p_0$ coordinates of $\boldsymbol{x}$ to observe, where $q_0 \le q$ and $p_0 \le p$. We use $[s]$ as a shorthand of the sequence $\{1, \ldots, s\}$ for any positive integer $s$. For a positive integer $d$ and $1 \le j \le d$, let $\boldsymbol{e}_j^{[d]}$ denote the $d$-dimensional column vector with the $j$th element being 1 and all other components being 0, and let $x_{t,j}$ and $y_{t,j}$ denote the $j$th coordinate of $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, respectively. For any $a \times b$ matrix $\boldsymbol{C}_t$ or $\boldsymbol{D}$, let $\boldsymbol{C}_{t,j:}$ or $\boldsymbol{D}_{j:}$ denote the $j$th row of $\boldsymbol{C}_t$ or $\boldsymbol{D}$, expressed as a row vector, and let $\boldsymbol{C}_{t,j:}^\mathsf{T}$ or $\boldsymbol{D}_{j:}^\mathsf{T}$ denote its transpose, $(\boldsymbol{C}_{t,j:})^\mathsf{T}$ or $(\boldsymbol{D}_{j:})^\mathsf{T}$, for simplicity. When expressing the $j$th row of the transpose $\boldsymbol{C}_t^\mathsf{T}$ of $\boldsymbol{C}_t$, we denote it as $(\boldsymbol{C}_t^\mathsf{T})_{j:}$.

### 3.1. Least Squares Ridge-like Regression

In contrast to (2), taking the loss function in (1) as the half of the $L_2$ norm, and given $T$, we consider minimizing the empirical squared loss at $t$, $\frac{1}{2}\|\boldsymbol{W}^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t\|^2$ for $1 \le t \le T$, whose expectation is the risk $R(\boldsymbol{W})$, given by (1) due to the i.i.d. assumption. Let $\boldsymbol{W}_t$ denote the weight matrix, also called the coefficient matrix, obtained from using an algorithm at $t$. Without full observations of outcomes and features, the key is to estimate the gradient of the squared loss at $\boldsymbol{W}_t$, $\boldsymbol{G}_t \triangleq \boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)^\mathsf{T}$, in an "unbiased" manner to be explained later. The coefficient matrix $\boldsymbol{W}_t$ is then updated to $\boldsymbol{W}_{t+1}$ using the gradient descent method for each $t = 1, \ldots, T$.

An online learning algorithm is proposed in Algorithm 1, where $\eta$ represents the step size and $B$ is the pre-specified tuning parameter as in (2).

When $q = q_0 = 1$, Algorithm 1 reduces to the AERR algorithm proposed by Hazan & Koren (2012), with the only difference being the sampling scheme in Line 5. Hence, Algorithm 1 virtually extends the AERR algorithm developed for univariate outcomes to accommodate multivariate outcomes. In Appendix A.1 we present an alternative multivariate AERR algorithm that applies the AERR procedure to each outcome variable separately. The superiority of the proposed Algorithm 1 to this multivariate AERR algorithm is demonstrated through simulations in Section 5.1.

The gradient $\breve{\boldsymbol{G}}_t$ constructed in Line 8 provides an unbiased

**Algorithm 1** Multivariate Least Squares Ridge Regression

**Initialize:** $B, \eta > 0$, $\boldsymbol{W}_1$ satisfying $0 < \|\boldsymbol{W}_1\|_F \leq B$, $q_0 \geq 1$, $p_0 \geq 2$
**for** $t = 1, \ldots, T$ **do**
  Choose $\{j_{t,r} : r \in [q_0]\}$ uniformly from $[q]$ without replacement
  $\tilde{\boldsymbol{y}}_t \leftarrow \frac{q}{q_0} \sum_{r=1}^{q_0} y_{t,j_{t,r}} \boldsymbol{e}_{j_{t,r}}^{[q]}$
  Choose $\{j_{t,r} : r \in [p_0 - 1]\}$ uniformly from $[p]$ without replacement
  $\tilde{\boldsymbol{x}}_t \leftarrow \frac{p}{p_0 - 1} \sum_{r=1}^{p_0 - 1} x_{t,j_{t,r}} \boldsymbol{e}_{j_{t,r}}^{[p]}$
  Pick an index $j_t$ with probability $\|\boldsymbol{W}_{t,j_t:}^{\mathsf{T}}\|^2 / \|\boldsymbol{W}_t\|_F^2$ from $[p]$
  $\breve{\boldsymbol{G}}_t \leftarrow \tilde{\boldsymbol{x}}_t (x_{t,j_t} \|\boldsymbol{W}_t\|_F^2 \boldsymbol{W}_{t,j_t:}^{\mathsf{T}} / \|\boldsymbol{W}_{t,j_t:}^{\mathsf{T}}\|^2 - \tilde{\boldsymbol{y}}_t)^{\mathsf{T}}$
  $\boldsymbol{V}_t \leftarrow \boldsymbol{W}_t - \eta \breve{\boldsymbol{G}}_t$
  $\boldsymbol{W}_{t+1} \leftarrow B \boldsymbol{V}_t / \max\{\|\boldsymbol{V}_t\|_F, B\}$
**end for**
**Return:** $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{W}_t$

estimate of $\boldsymbol{G}_t$ conditional on $\boldsymbol{W}_t$, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$. The proof of this property is provided in (24) of Appendix C. This property contributes to the following theorem about Algorithm 1.

**Theorem 3.1.** *Assume that (4) holds. Let* $\hat{\boldsymbol{W}} \triangleq \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{W}_t$ *denote the output of Algorithm 1. Then* $\|\hat{\boldsymbol{W}}\|_F \leq B$, *and furthermore,*

*(a). for any* $p \times q$ *matrix* $\boldsymbol{W}^*$ *with* $\|\boldsymbol{W}^*\|_F \leq B$,

$$\mathbb{E}\{R(\hat{\boldsymbol{W}})\} \leq R(\boldsymbol{W}^*) + \frac{2B^2}{\eta T} + \frac{\eta B_x^2 p}{p_0 - 1} \left( \frac{q B_y^2}{q_0} + B^2 B_x^2 \right);$$

*(b). with the step size in Algorithm 1 given as*

$$\eta = \sqrt{\frac{2B^2(p_0 - 1)}{T B_x^2 p (B^2 B_x^2 + q B_y^2 / q_0)}}, \tag{5}$$

*we have that*

$$\mathbb{E}\{R(\hat{\boldsymbol{W}})\} \leq R(\boldsymbol{W}^*) + 4\tilde{B}^2 \sqrt{\frac{pq}{T(p_0 - 1)q_0}}, \tag{6}$$

*where* $\tilde{B} \triangleq \max\{B B_x, B_y\}$.

The result (6) implies that to learn $\hat{\boldsymbol{W}}$ with $\|\hat{\boldsymbol{W}}\|_F \leq B$, such that $R(\hat{\boldsymbol{W}}) - R(\boldsymbol{W}^*) \leq \epsilon$ in expectation for any $\|\boldsymbol{W}^*\|_F \leq B$, at least $O\left(\frac{pq}{(p_0 - 1)q_0 \epsilon^2}\right)$ examples are required. This characterization of sample complexity extends the counterpart of Theorem 3.1 in Hazan & Koren (2012).

**Algorithm 2** Multivariate Least Squares Lasso Regression

**Initialize:** $\lambda, \eta_t > 0$, $\boldsymbol{W}_1$ satisfying $0 < \|\boldsymbol{W}_1\|_F \leq B$, $q_0 \geq 1$, $p_0 \geq 2$
**for** $t = 1, \ldots, T$ **do**
  Construct $\breve{\boldsymbol{G}}_t$ as in Line 8 of Algorithm 1
  $\boldsymbol{V}_t \leftarrow \boldsymbol{W}_t - \eta_t \breve{\boldsymbol{G}}_t$
  $\boldsymbol{W}_{t+1}^* \leftarrow \underset{\boldsymbol{W}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{W} - \boldsymbol{V}_t\|_F^2 + \eta_t P_\lambda(\boldsymbol{W}) \right\}$
  $\boldsymbol{W}_{t+1} \leftarrow B \boldsymbol{W}_{t+1}^* / \max\{\|\boldsymbol{W}_{t+1}^*\|_F, B\}$
**end for**
**Return:** $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{W}_t$

Specifically, if $q = q_0$, the second term on the right hand side of (6) reduces to the counterpart in Hazan & Koren (2012), except for a small constant difference induced from using a different strategy to sample the observable indices. In addition, Markov's inequality yields that for any $\epsilon, \delta > 0$

$$\mathbb{P}\left( R(\hat{\boldsymbol{W}}) - \inf_{\|\boldsymbol{W}\|_F \leq B} R(\boldsymbol{W}) \geq \epsilon \right) \leq \delta$$

if $T \geq \frac{16 \tilde{B}^4 pq}{(p_0 - 1)q_0} \epsilon^{-2} \delta^{-2}$, indicating that Algorithm 1 is a PAC learning algorithm (Mohri et al., 2018, Chapter 2).

### 3.2. Least Squares Lasso-type Regression

In contrast to (3), we extend the development in Section 3.1 to include sparsity-inducing penalties. For any $p \times q$ matrix $\boldsymbol{W}$ and for $l = 1, \ldots, p$, we consider the penalty function:

$$P_\lambda(\boldsymbol{W}) = \lambda_1 \sum_{l=1}^{p} \|\boldsymbol{W}_{l:}^{\mathsf{T}}\| + \lambda_2 \sum_{l=1}^{p} \|\boldsymbol{W}_{l:}^{\mathsf{T}}\|_1, \tag{7}$$

where $\lambda = (\lambda_1, \lambda_2)^{\mathsf{T}}$ includes tuning parameters $\lambda_1$ and $\lambda_2$. When $\lambda_1 = 0$, the penalty function (7) reduces to Lasso, which shrinks each element of $\boldsymbol{W}$ towards 0, aiding in the detection of active features for each coordinate of the outcome. When $\lambda_2 = 0$, (7) represents the group Lasso, which shrinks each row of $\boldsymbol{W}$ towards $\boldsymbol{0}$, and thus, facilitating the detection of active features for the entire outcome. The general sparse group lasso penalty (7) achieves both levels of selection simultaneously.

Let $\{\eta_t : t = 1, 2, \ldots\}$ denote a collection of step sizes. To address the learning of multivariate outcomes with limited observations, we introduce Algorithm 2, which modifies the FOBOS method proposed by Duchi & Singer (2009). In this algorithm, the key step in Line 5, which is indeed the proximal mapping of $\eta_t P_\lambda(\boldsymbol{W})$ (Beck, 2017, Chaper 7), produces a matrix close to $\boldsymbol{V}_t$, acquired through the unconstrained gradient descent step in Line 4. This process also attains certain sparsity due to the penalty term. The regularization parameter $B > 0$ is defined in (3).

One might inquire about the existence of a more efficient method for the optimization problem in Line 5 of Algorithm

2. The following proposition validates the existence of a closed-form solution, with its proof deferred to Section C.3 of the supplementary material. We adopt the convention that $0/0 = 0$. Define the soft thresholding operator as $S_\mu(x) = (|x| - \mu)_+ x/|x|$ for $x \in \mathbb{R}$ and $\mu \geq 0$, where $(u)_+ \triangleq \max\{u, 0\}$ for any $u \in \mathbb{R}$. For a vector $\boldsymbol{x} = (x_1, \ldots, x_d)^\mathsf{T}$, let $S_\mu(\boldsymbol{x})$ denote the vector $(S_\mu(x_1), \ldots, S_\mu(x_d))^\mathsf{T}$.

**Proposition 3.2.** *The $j$th row of $\boldsymbol{W}^*_{t+1}$ in Algorithm 2 can be represented as:*

$$\boldsymbol{W}^*_{t+1,j:} = (\|S_{\eta_t\lambda_2}(\boldsymbol{V}_{t,j:})\| - \eta_t\lambda_1)_+ \cdot \frac{S_{\eta_t\lambda_2}(\boldsymbol{V}_{t,j:})}{\|S_{\eta_t\lambda_2}(\boldsymbol{V}_{t,j:})\|}$$

*for $j = 1, \ldots, p$.*

**Theorem 3.3.** *Assume that (4) holds. Let $\hat{\boldsymbol{W}}$ denote the output of Algorithm 2. Then $\|\hat{\boldsymbol{W}}\|_F \leq B$, and futhermore,*

*(a). for any $p \times q$ matrix $\boldsymbol{W}^*$ with $\|\boldsymbol{W}^*\|_F \leq B$,*

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \leq R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + \Delta_1,$$

*where*

$$\Delta_1 \triangleq \frac{pB_x^2\bar{\eta}}{p_0 - 1}\left(\frac{qB_y^2}{q_0} + B^2B_x^2\right) + \frac{2B^2}{\eta_T T} + \frac{1}{T}P_\lambda(\boldsymbol{W}_1)$$

*with $\bar{\eta} = \frac{1}{T}\sum_{t=1}^T \eta_t$;*

*(b). with $\eta_t = \eta$ for all $t = 1, \ldots, T$ and $\eta$ given by (5),*

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \leq R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + \Delta_1', \tag{8}$$

*where*

$$\Delta_1' \triangleq 4\tilde{B}^2\sqrt{\frac{pq}{T(p_0 - 1)q_0}} + \frac{1}{T}P_\lambda(\boldsymbol{W}_1)$$

*and $\tilde{B} \triangleq \max\{BB_x, B_y\}$.*

*(c). with $\{\eta_t : t = 1, 2, \ldots\}$ satisfying that*

$$\eta_T T \to \infty \text{ and } \frac{1}{T}\sum_{t=1}^T \eta_t \to 0 \text{ as } T \to \infty,$$

*we have that*

$$\lim_{T\to\infty} \mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} = R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*),$$

*where $\boldsymbol{W}_0^*$ is optimal in the sense that for any given $\lambda \geq 0$,*

$$R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*) = \inf_{\|\boldsymbol{W}\|_F \leq B}\{R(\boldsymbol{W}) + P_\lambda(\boldsymbol{W})\}.$$

Theorem 3.3 encompasses Theorem 3.1 as a special case when $\lambda_1 = \lambda_2 = 0$. A typical choice for step sizes involves setting $\eta_t \propto 1/\sqrt{t}$ for $t = 1, 2, \ldots$, which decreases as $t$ increases. Such step sizes, similar to (8) with a fixed step size, yields an $O(\sqrt{1/T})$ bound as $T \to \infty$ (Zinkevich, 2003).

**Algorithm 3** Multivariate Least Absolute Deviations Ridge Regression

**Initialize:** $B, \eta > 0$, $\boldsymbol{W}_1$ satisfying $\|\boldsymbol{W}_1\|_F \leq B$, $q_0 \geq 1$, $p_0 \geq 1$
**for** $t = 1, \ldots, T$ **do**
  Choose $\{j_{t,r} : r \in [p_0]\}$ uniformly from $[p]$ without replacement
  $\tilde{\boldsymbol{x}}_t \leftarrow \sum_{r=1}^{p_0} x_{t,j_{t,r}} \boldsymbol{e}^{[p]}_{j_{t,r}}$
  Choose $\mathcal{O}_t \triangleq \{j_{t,r} : r \in [q_0]\}$ uniformly from $[q]$ without replacement
  $\boldsymbol{\phi}_{t,\mathcal{O}_t} \leftarrow$ any element from the set $\partial\|\boldsymbol{W}^\mathsf{T}_{t,:\mathcal{O}_t}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1$
  $\breve{\boldsymbol{G}}_t \leftarrow \frac{q}{q_0}\tilde{\boldsymbol{x}}_t\boldsymbol{\phi}_t^\mathsf{T}$, where $\boldsymbol{\phi}_t \in \mathbb{R}^q$ with $\mathcal{O}_t$-elements being $\boldsymbol{\phi}_{t,\mathcal{O}_t}$ and the remaining being 0
  $\boldsymbol{V}_t \leftarrow \boldsymbol{W}_t - \eta\breve{\boldsymbol{G}}_t$
  $\boldsymbol{W}_{t+1} \leftarrow B\boldsymbol{V}_t/\max\{\|\boldsymbol{V}_t\|_F, B\}$
**end for**
**Return:** $\frac{1}{T}\sum_{t=1}^T \boldsymbol{W}_t$

## 4. Algorithms for Least Absolute Deviations

### 4.1. Least Absolute Deviations Ridge-like Regression

In this subsection, we propose an algorithm to solve the problem (2) using the $L_1$ loss function, with an extension feasible for accommodating general Lipschitz convex loss functions. Here, we use notations analogous to those in Algorithm 1. For a $d \times 1$ vector $\boldsymbol{u} = (u_1, \ldots, u_d)^\mathsf{T}$, with $d$ being a positive integer, let $\partial\|\boldsymbol{u}\|_1$ denote the subgradient of function $\|\cdot\|_1$ at $\boldsymbol{u}$, with the $j$th component denoted $\partial|u_j|$ for $j = 1, \ldots, d$, where $\partial|u_j|$ equals the sign of $u_j$ when $u_j \neq 0$ and $\partial|u_j| = [-1, 1]$ when $u_j = 0$. For a vector $\boldsymbol{v} \in \mathcal{R}^q$ and a subset $\mathcal{O} \subset \{1, \ldots, q\}$, let $\boldsymbol{v}_\mathcal{O}$ denote the subvector of $\boldsymbol{v}$ restricted on the indices set $\mathcal{O}$. For any $p \times q$ matrix $\boldsymbol{W}_t$, $\boldsymbol{W}_{t,:\mathcal{O}}$ represents the $p \times |\mathcal{O}|$ submatrix with columns indices $\mathcal{O}$, where $|\mathcal{O}|$ is the cardinality of $\mathcal{O}$.

The crucial aspect of Algorithm 1 lies in the construction of an unbiased estimate of the gradient of the $L_2$ loss. However, this approach becomes infeasible now when dealing with the absolute value loss function that is non-differentiable. Instead, we resort to subgradient methods and carefully manage the resulting error bound to ensure PAC guarantees when there are no missing input features (i.e., $p_0 = p$), as shown in Theorem 4.1. It is important to note that introducing multi-dimensional outcomes poses challenges when employing the absolute value loss function. Merely computing the subgradient of the absolute value loss function at $\boldsymbol{W}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \tilde{\boldsymbol{y}}_t$ would lead to an unsatisfactory error bound due to substantial bias, where $\tilde{\boldsymbol{x}}_t$ and $\tilde{\boldsymbol{y}}_t$ are constructed similarly to those in Algorithm 1. This issue, however, diminishes when dealing with univariate outcome.

**Theorem 4.1.** *Assume that (4) holds. Let $\hat{\boldsymbol{W}}$ denote the*

**Algorithm 4** Multivariate Least Absolute Deviations Lasso Regression

**Initialize:** $B, \eta_t > 0$, $\boldsymbol{W}_1$ satisfying $\|\boldsymbol{W}_1\|_F \le B$, $q_0 \ge 1, p_0 \ge 1$
**for** $t = 1, \dots, T$ **do**
  Construct $\breve{\boldsymbol{G}}_t$ as in Line 8 of Algorithm 3
  $\boldsymbol{V}_t \leftarrow \boldsymbol{W}_t - \eta_t \breve{\boldsymbol{G}}_t$
  $\boldsymbol{W}^*_{t+1} \leftarrow \operatorname{argmin}_{\boldsymbol{W}} \left\{ \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{V}_t\|_F^2 + \eta_t P_\lambda(\boldsymbol{W}) \right\}$
  $\boldsymbol{W}_{t+1} \leftarrow B\boldsymbol{W}^*_{t+1} / \max\{\|\boldsymbol{W}^*_{t+1}\|_F, B\}$
**end for**
**Return:** $\frac{1}{T}\sum_{t=1}^T \boldsymbol{W}_t$

*output of Algorithm 3. Then* $\|\hat{\boldsymbol{W}}\|_F \le B$, *and futhermore,*

*(a). for any* $p \times q$ *matrix* $\boldsymbol{W}^*$ *with* $\|\boldsymbol{W}^*\|_F \le B$,

$$\mathbb{E}\{R(\hat{\boldsymbol{W}})\} \le R(\boldsymbol{W}^*) + \frac{2B^2}{\eta T} + \frac{q^2 p_0 B_x^2 \eta}{2pq_0} + 2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}};$$

*(b). with* $\eta$ *in Algorithm 3 set as* $\eta = \frac{2B}{B_x q}\sqrt{\frac{pq_0}{p_0 T}}$,

$$\mathbb{E}\{R(\hat{\boldsymbol{W}})\} \le R(\boldsymbol{W}^*) + 2BB_x\sqrt{\frac{q^2 p_0}{pq_0 T}} + 2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}}. \quad (9)$$

Theorem 4.1 extends the results in Bullins et al. (2016, Theorem 8) from the case with $q = q_0 = 1$ to accommodate multivariate outcomes with $q \ge 1$. The expression (9) provides an upper bound for the expected risk of the algorithm output. If $p = p_0$, the last term in (9) becomes exactly 0, and thus, Algorithm 3 is a PAC-learning algorithm, following from similar arguments after Theorem 3.1. On the contrary, when $p_0 < p$, unlike the last term in (6) that approaches 0 as $T \to \infty$, the last term in (9) remains a constant irrespective of $T$. Indeed, Bullins et al. (2016, Corollary 4) showed that for the case with $q_0 = q = 1$, no PAC algorithm exists if $p_0 < p$.

### 4.2. Least Absolute Deviations Lasso-type Regression

Analogous to Section 3.2, we now study least square deviations regression with sparsity-inducing penalty (7) included. In the same spirit of Algorithm 2, we propose Algorithm 4, whose theoretical guarantee is presented as follows.

**Theorem 4.2.** *Assume that (4) holds. Let* $\hat{\boldsymbol{W}}$ *denote the output of Algorithm 4. Then* $\|\hat{\boldsymbol{W}}\|_F \le B$, *and furthermore,*

*(a). for an any* $p \times q$ *matrix* $\boldsymbol{W}^*$ *with* $\|\boldsymbol{W}^*\|_F \le B$,

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \le R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + \Delta_2,$$

*where*

$$\Delta_2 \triangleq \frac{2B^2}{\eta_T T} + \frac{q^2 p_0 B_x^2 \bar{\eta}}{2pq_0} + 2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}} + \frac{P_\lambda(\boldsymbol{W}_1)}{T}$$

*with* $\bar{\eta} = \frac{1}{T}\sum_{t=1}^T \eta_t$;

*(b). with* $\eta_t$ *in Algorithm 4 set as* $\eta_t = \frac{2B}{B_x q}\sqrt{\frac{pq_0}{p_0 T}}$ *for all* $t = 1, \dots, T$,

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \le R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + \Delta_2',$$

*where*

$$\Delta_2' \triangleq 2BB_x\sqrt{\frac{q^2 p_0}{pq_0 T}} + 2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}} + \frac{P_\lambda(\boldsymbol{W}_1)}{T};$$

*(c). with* $\{\eta_t : t = 1, 2, \dots\}$ *satisfying that*

$$\eta_T T \to \infty \text{ and } \frac{1}{T}\sum_{t=1}^T \eta_t \to 0 \text{ as } T \to \infty,$$

*we have that*

$$\limsup_{T\to\infty} \mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \le R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + 2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right).$$

Theorem 4.2 further extends Theorem 4.1 by incorporating an additional penalization and decreasing step sizes. Assuming $\boldsymbol{W}_0^*$ is optimal as in Theorem 4.1, it is straightforward to show that

$$\liminf_{T\to\infty} \mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \ge R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*).$$

Unlike Theorem 3.3(d) which delineates the limit of $\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\}$ for the output of Algorithm 2, here, for the output of Algorithm 4, we can only characterize the inferior and superior limits of $\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\}$, which are bounded between $R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*)$ and $R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*) + 2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}}$. The term $2\sqrt{q}BB_x\sqrt{1-\frac{p_0}{p}}$ indicates that if $p_0 < p$, there may exist a gap between the output of the algorithm and the optimal weight matrix $\boldsymbol{W}_0^*$ even as $T \to \infty$.

## 5. Experiments

### 5.1. Synthetic Data: Merit of Multivariate Algorithms

The algorithms proposed in Sections 3 and 4 consider all observed outcomes simultaneously. While one might question the necessity of such an approach, as existing algorithms for univariate observable outcomes can be extended directly, we

demonstrate the superiority of Algorithm 1 over the "multivariate AERR" using various synthetic datasets. The AERR algorithm, proposed by Hazan & Koren (2012), is designed for one dimensional outcomes and incomplete features. In essence, we execute the AERR procedure separately for each observed coordinate of outcomes. The detailed pseudo code of "multivariate AERR" is presented in Section A.1.

Set $q = 5$, $q_0 = 2$, $p = 20$, and $p_0 = 10$. Let $\boldsymbol{W}_0$ be a $p \times q$ matrix whose coordinates are chosen from the uniform distribution over $\{1, -1, 2, -2\}$. For $i = 1, \ldots, T$, we generate each feature vector $\boldsymbol{x}_i \in \mathbb{R}^p$ independently from a centered multivariate normal distribution having variance matrix with $(j, j')$ element being $0.5^{|j-j'|}$. Outcomes are then generated from the linear model:

$$\boldsymbol{y}_i = \boldsymbol{W}_0^{\mathsf{T}} \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i \quad \text{for} \quad i = 1, \ldots, T, \tag{10}$$

where $\boldsymbol{y}_i \in \mathbb{R}^q$, and the random error $\boldsymbol{\varepsilon}_i \in \mathbb{R}^q$ is sampled from the normal distribution having mean $\mathbf{0}$ and variance matrix with the $(j, j')$ element being $\sigma_\varepsilon^2 \cdot 0.1^{|j-j'|}$.

We examine various settings for the sample size $T$ (10000, 20000 or 50000) and the noise level $\sigma_\varepsilon^2$ (5 or 10). In each setting, the entire procedure is repeated 300 times to compare the performance of three methods by presenting the associated average results: Algorithm 1 (abbreviated as LSR), multivariate AERR with $p_0 = 10$ (abbreviated as AERR1), and multivariate AERR with $p_0$ redefined as 12 (abbreviated as AERR2). A summary of the observed features is displayed in Table 2. Specifically, since AERR involves sampling observation indices with replacement, differing from Line 6 in Algorithm 1, there are instances where the total number of observations may be smaller than that of Algorithm 1. To address this, we set a larger $q_0$ to ensure that the number of observations is slightly larger than that of Algorithm 1. The comparison between LSR and AERR1 is based on the same sample size $T$, whereas the comparison between LSR and AERR2 focuses on a similar total number of observations. In implementing all methods, $B$ is set to be 100.

According to Hazan & Koren (2012, Theorem 3.1), the optimal constant step size for AERR is of order $O\left(\sqrt{\frac{p_0-1}{2pT_*}}\right)$, where $T_* = q_0 T/q$ represents an effective sample size, given that each coordinate of outcomes is observed with the probability of $q_0/q$. We set the step size to be $\frac{1}{9}\sqrt{\frac{p_0-1}{2pT_*}}$. By Theorems 3.1, the optimal step size for Algorithm 1 is $O\left(\sqrt{\frac{2(p_0-1)}{Tp(1+q/q_0)}}\right)$. Hence, the step size for LSR is set to be $\frac{1}{9}\sqrt{\frac{2(p_0-1)}{Tp(1+q/q_0)}}$.

For additional comparisons, we employ two imputation methods utilizing the R package 'mice'. The first method, referred to as 'Imp1', involves imputation on the entire dataset using the 'mice' function. Missing indices are generated using LSR (Algorithm 1), followed by using the multivariate linear regression model to estimate coefficients. In contrast, the second method, denoted 'Imp2', employs mean imputation for missing feature attributes, followed by employing 'mice' to impute missing outcomes.

*Table 2.* Synthetic data: Mean of the observed features attributes across 300 replicates.

| | $T = 10^4$ | $T = 2 \times 10^4$ | $T = 5 \times 10^4$ |
|---|---|---|---|
| LSR | 95500.26 | 190999.3 | 477498.2 |
| AERR1 | 86237.33 | 172480.4 | 431179.1 |
| AERR2 | 97327.83 | 194657.8 | 486670.9 |

The performance of each method is evaluated by the mean prediction error: $\frac{1}{T'}\sum_{i=1}^{T'} \|\boldsymbol{y}_i^* - \hat{\boldsymbol{W}}^{\mathsf{T}} \boldsymbol{x}_i^*\|^2$, where $\boldsymbol{y}_i^*$ and $\boldsymbol{x}_i^*$ represent new, fully observed samples generated independently and identically to the training sample, and $\hat{\boldsymbol{W}}$ denotes the output obtained from method. We set $T' = 5000$ and display the results in Table 3. As anticipated, a larger $q_0$ makes AERR2 perform better than AERR1, although the improvement is marginal. Additionally, the prediction error decreases as the sample size increases or the noise level decreases. Across all settings, the proposed LSR method significantly outperforms AERR1 and even AERR2, which requires a larger $q_0$ for each sample. This advantage of LSR over AERR may stem from Line 8 of Algorithm 1, where the gradient computation is not only more efficient compared to Lines 7-9 of Algorithm 5 but also more effective in terms of optimization. Finally, Imp1 greatly outperforms all other methods, whereas Imp2 performs worse than our method. However, it is worth noting that comparing Imp1 and Imp2 with other three methods is not fair, as the former methods are batch mode, utilizing the entire dataset at the cost of increased computation, whereas the latter methods use the data sequentially with a smaller size of data at each time point.

### 5.2. Synthetic Data: Robustness

We conduct simulations to compare Algorithms 1 and 3 in this subsection. The data generation setting is analogous to that in Section 5.1, with $\sigma_\varepsilon^2$ now fixed at 5 and $p_0$ allowed to vary. While Algorithm 3 (abbreviated as LADR) cannot produce arbitrarily precise prediction as $T \to \infty$, as per Theorem 4.1, in certain situations with outliers, LADR may outperform Algorithm 1. Specifically, we introduce 20% of training samples as outliers, with their corresponding random errors generated independently from the standard Cauchy distribution with center 0 and scale 1.

We consider different sample sizes for $T$ (10000,20000 or 50000) and values for $p_0$ (10,15 or 20). We compare LADR to Algorithm 1 (abbreviated as LSR) and Algorithm

*Table 3.* Experiment results: Mean (and standard variation) of mean prediction errors over 300 replicates for LSR, AERR1, AERR2, Imp1, and Imp2.

| $\sigma_\varepsilon^2$ | | $T = 1 \times 10^4$ | $T = 2 \times 10^4$ | $T = 5 \times 10^4$ |
|---|---|---|---|---|
| 5 | LSR | 25.08 (1.65) | 20.38 (1.13) | 16.31 (0.60) |
| | AERR1 | 37.41 (4.18) | 30.18 (3.14) | 22.19 (1.49) |
| | AERR2 | 36.77 (7.13) | 28.99 (3.06) | 21.52 (2.04) |
| | IMP1 | 14.18 (0.24) | 13.88 (0.21) | 13.68 (0.17) |
| | IMP2 | 28.91 (3.21) | 28.82 (3.29) | 28.62 (3.34) |
| 10 | LSR | 37.63 (1.69) | 32.95 (1.16) | 28.84 (0.66) |
| | AERR1 | 49.54 (5.41) | 42.21 (2.60) | 34.84 (1.80) |
| | AERR2 | 48.69 (3.86) | 41.42 (2.78) | 33.86 (1.37) |
| | IMP1 | 26.46 (0.31) | 26.09 (0.26) | 25.86 (0.27) |
| | IMP2 | 41.53 (3.31) | 41.25 (3.28) | 41.00 (3.36) |

*Table 4.* Experiment results: Mean (and standard variation) of mean prediction errors over 300 replicates for LADR, LSR and LSR0.

| $T(10^4)$ | | $p_0 = 10$ | $p_0 = 15$ | $p_0 = 20$ |
|---|---|---|---|---|
| 1 | LSR | 79.66 (73.76) | 66.53 (60.23) | 51.25 (38.94) |
| | LADR | 29.96 (3.27) | 17.14 (0.64) | 12.70 (0.12) |
| | LSR0 | 25.19 (1.68) | 21.36 (1.08) | 18.95 (0.87) |
| 2 | LSR | 77.39 (62.29) | 53.75 (37.55) | 35.85 (21.02) |
| | LADR | 29.56 (3.49) | 16.72 (0.60) | 12.59 (0.11) |
| | LSR0 | 20.41 (1.05) | 17.62 (0.73) | 16.13 (0.53) |
| 5 | LSR | 52.76 (30.82) | 38.81 (19.24) | 29.25 (12.15) |
| | LADR | 29.89 (3.63) | 16.60 (0.60) | 12.53 (0.11) |
| | LSR0 | 16.32 (0.57) | 14.84 (0.38) | 14.14 (0.28) |

1 under non-contaminated samples (abbreviated as LSR0) with the same sample size. The average results for 300 repeated implementations are displayed in Table 4. First, LADR significantly outperforms LSR under all settings, showing the robustness of Algorithm 3 in the presence of outliers. Secondly, in some situations, especially when $p_0$ is large and $T$ is small, LADR is even more accurate and stable than LSR0. Thirdly, increasing $p_0$ improves the prediction accuracy for all three methods. On the other hand, increasing $T$ decreases the prediction error for LSR and LSR0 only. This suggests that Algorithm 3 can be more efficient than Algorithm 1 under certain finite sample cases, although theoretically, it may not be as accurate as Algorithm 1 when $T \to \infty$. The comparison of Algorithms 2 and 4 follows a similar pattern and is deferred to Section A.2.

### 5.3. Synthetic Data: Penalization

In this subsection, we present simulation results for Algorithm 2 and compare it to Algorithm 1. We consider a setting with $q = 5$, $q_0 = 2$, and $p = 50$. The elements of the first 5 rows of $\boldsymbol{W}_0$ in model (10) are generated from $\{0, 1, -1, 2, -2\}$ with equal probabilities of 0.2 each, while the remaining elements of other 45 rows are all set to 0. The distributions of $\boldsymbol{x}_i$ and $\boldsymbol{\varepsilon}_i$ are identical to those in Section 5.1, except that the noise level $\sigma_\varepsilon^2$ is fixed to be 1. The two tuning parameters in (7), $\lambda_1$ and $\lambda_2$, are set to 0.1 and 0.001, respectively. The choices of $B$ and step size $\eta$ are the same as those in Section 5.1.

*Table 5.* Experiment results: Mean (and standard variation) of PEP, PE and PE0 over 300 replicates.

| $p_0$ | | $T = 4 \times 10^4$ | $T = 6 \times 10^4$ | $T = 8 \times 10^4$ |
|---|---|---|---|---|
| 20 | PEP | 35.10 (12.87) | 17.77 (4.45) | 12.26 (2.32) |
| | PE | 28.15 (11.53) | 12.81 (3.65) | 8.21 (1.77) |
| | PE0 | 43.27 (17.27) | 18.14 (5.26) | 11.61 (2.87) |
| 40 | PEP | 8.92 (1.16) | 7.04 (0.72) | 6.34 (0.51) |
| | PE | 5.73 (0.79) | 4.33 (0.41) | 3.81 (0.27) |
| | PE0 | 7.21 (1.02) | 5.16 (0.56) | 4.38 (0.35) |

We calculate the mean prediction error with penalty (abbreviated as PEP) for $T' = 5000$ new, fully observed samples: $\frac{1}{T'} \sum_{i=1}^{T'} \|\boldsymbol{y}_i^* - \hat{\boldsymbol{W}}_2^\mathsf{T} \boldsymbol{x}_i^*\|^2 + P_\lambda(\hat{\boldsymbol{W}}_2)$, where $\hat{\boldsymbol{W}}_2$ represents the output of Algorithm 2. In addition, we compare the mean prediction error of Algorithm 2 (abbreviated as PE) and Algorithm 1 (abbrevaited as PE0): $\frac{1}{T'} \sum_{i=1}^{T'} \|\boldsymbol{y}_i^* - \hat{\boldsymbol{W}}_l^\mathsf{T} \boldsymbol{x}_i^*\|^2$, where $l = 1$ for PE0 and $l = 2$ for PE.

Table 5 displays average values for PEP, PE and PE0 under various values of $p_0$ and $T$. It is evident that all three criteria decrease as $T$ increases. A smaller $p_0$ requires a larger sample size to achieve comparable prediction accuracy. The comparison between PE and PE0 indicates that Algorithm 2 may yield better predictions than Algorithm 1, especially when $p$ is large and $p_0$ is small. The advantage of penalization for the least absolute deviation loss is shown in Section A.2.

### 5.4. Yeast Cell Data

We apply the proposed method to the yeast cell dataset, available from R package "*spls*". The objective is to explore the influence of transcription factors (TFs) on the regulation of the yeast cell cycle. The outcomes represent gene expression measurements of 542 genes related to the cell cycle. A total of 18 mRNA levels are measured every 7 minutes over a period of 119 minutes. The features include the binding information for 106 transcription factors for these genes. Using the notation in Section 3, we have $p = 106$, $q = 18$, and $T = 542$. For futher details, refer to Chun & Keleş (2010).

We centralize the outcomes at each time point by subtracting the sample average from individual measurements. Each TFs feature is standardized such that the empirical mean and variance are 0 and 1, respectively. We randomly split the entire sample into training data and test data using 10 fold cross validation. For each split, Algorithm 1 is applied to training data, with $B = 2$, and $\eta$ is chosen similar to that in Section 5.1. The prediction error is then computed using test data, and the mean prediction error (MPE) is computed for those 10 prediction error values obtained from the 10 splits. To assess the performance of the proposed Algorithm 1 under different values of $p_0$ and $q_0$, we consider settings with $p_0 \in \{5, 10, 15, 20\}$ and $q_0 \in \{5, 10, 15\}$ and plot the mean of MPE values over 50 repetitions of the random split cross-validation procedure in Figure 1. For comparison, we also report results obtained using the "multivariate AERR" method. Analogous to the results in Section 5.1, it is clearly that Algorithm 1 (LSR) consistently outperforms the multivariate AERR across various settings of $(p_0, q_0)$. Both methods produce smaller prediction error as $p_0$ or $q_0$ increases.

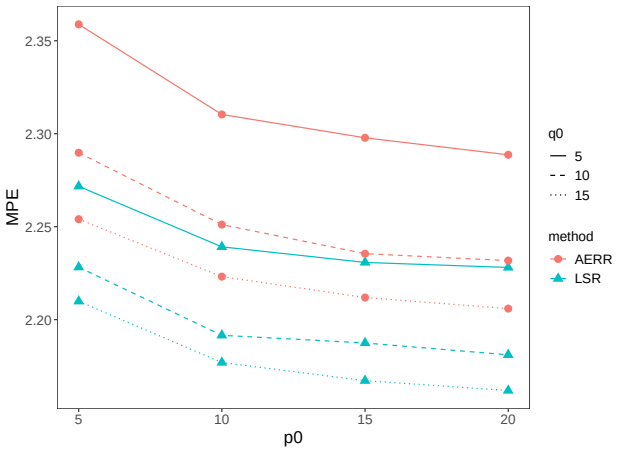


*Figure 1.* Application results for yeast cell data: The mean prediction error for $p_0 = 5$, 10, 15, 20, and $q_0 = 5$ (solid line), 10 (dash line), 15 (dotted line), using Algorithm 1 (LSR, blue triangle) or the "multivariate AERR" (red dot).

## 6. Discussion

Within the framework of multivariate linear regression, we explore a scenario where learners are constrained to observe only a limited number of attributes for both outcomes and features. We introduce efficient and easily implementable algorithms tailored to various loss functions and regularization techniques. Our research establishes the PAC property for the least squares loss function, while highlighting the infeasibility of achieving PAC for algorithms employing the least absolute value loss function unless the features are fully observed. This work significantly expands the analytical scope beyond existing investigations on univariate regression with restricted feature observation. Extensive experiments underscore the superiority of our proposed method over the naive method that applies univariate outcome algorithms to individual components of outcomes.

In line with common practice in the literature, our development assumes that $\{(\boldsymbol{x}_t, \boldsymbol{y}_t) : t = 1, 2, \ldots\}$ are independent and identically distributed (iid). Departure from this iid assumption may affect the applicability of our proposed methods. When the assumption is not strictly met but remains reasonably close to the iid scenario, our learning results can still provide reasonably good approximations to the underlying truth. However, if the iid assumption is deemed completely implausible, it is crucial to exercise extra caution when interpreting the results of our learning algorithms.

While all algorithms and associated theorems fall within the scope of online learning, inspired by the work of Hazan & Koren (2012), we do not necessarily view online learning as indispensable for managing missing data in our context. Alternative approaches, such as batch mode algorithms, can also be developed by modifying our current development.

An interesting future exploration involves bridging the gap between $L_1$ regression with complete observation and incomplete observation. Although building a PAC learning algorithm within the considered framework is deemed impossible (Bullins et al., 2016), the gap may be asymptotically eliminated if a certain proportion of the entire sample is fully observed. If we can select the number of instances for their full information access, a natural question arises: how many queries are needed to achieve a good approximation of the optimal solution across the entire dataset?

Our development utilizes multiple linear regression models, known for their simplicity and transparent interpretations. It is interesting to generalize our methods to accommodate other models that facilitate complex relationships between outcomes and features. For example, exploring the integration of achine learning methods such as multi-outcome deep learning or tree ensemble methods would be useful. We anticipate that establishing theoretical results may become more challenging than our derivations here.

## Acknowledgements

The authors thank the Program Chair and all the reviewers for their helpful comments on the initial manuscript. Grace Y. Yi is Canada Research Chair in Data Science (Tier 1). Her research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs program.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References


Baltagi, B. H. and Baltagi, B. H. *Econometric Analysis of Panel Data*. Springer, 2008.

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

Ben-David, S. and Dichterman, E. Learning with restricted focus of attention. In *Proceedings of the 6th Annual Conference on Computational Learning Theory*, pp. 287–296, 1993.

Ben-David, S. and Dichterman, E. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.

Breiman, L. and Friedman, J. H. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society, Series B*, 59(1):3–54, 1997.

Bullins, B., Hazan, E., and Koren, T. The limits of learning with missing data. *Advances in Neural Information Processing Systems*, 29, 2016.

Cesa-Bianchi, N., Shalev-Shwartz, S., and Shamir, O. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 12(10), 2011.

Chen, B., Yi, G. Y., and Cook, R. J. Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*, 105(489):336–353, 2010.

Chen, Q., Ibrahim, J. G., Chen, M.-H., and Senchaudhuri, P. Theory and inference for regression models with missing responses and covariates. *Journal of Multivariate Analysis*, 99(6):1302–1331, 2008.

Chun, H. and Keleş, S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, 72(1):3–25, 2010.

Dekel, O. and Shamir, O. Learning to classify with missing and corrupted features. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 216–223, 2008.

Diggle, P. *Analysis of Longitudinal Data*. Oxford university press, 2002.

Duchi, J. and Singer, Y. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research*, 10:2899–2934, 2009.

Foster, D., Kale, S., and Karloff, H. Online sparse linear regression. In *Conference on Learning Theory*, pp. 960–970, 2016.

Gong, Y., Li, Z., Liu, W., Lu, X., Liu, X., Tsang, I. W., and Yin, Y. Missingness-pattern-adaptive learning with incomplete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11053–11066, 2023.

Han, Y., Sun, G., Shen, Y., and Zhang, X. Multi-label learning with highly incomplete data via collaborative embedding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1494–1503, New York, NY, USA, 2018. Association for Computing Machinery.

Hao, X., Huang, J., Qin, F., and Zheng, X. Multi-label learning with missing features and labels and its application to text categorization. *Intelligent Systems with Applications*, 14:200086, 2022.

Hazan, E. and Koren, T. Linear regression with limited observation. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pp. 1865–1872, 2012.

Hazan, E., Livni, R., and Mansour, Y. Classification with low rank and missing data. In *International Conference on Machine Learning*, pp. 257–266, 2015.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Horváth, L. and Kokoszka, P. *Inference for Functional Data with Applications*. Springer Science & Business Media, 2012.

Ibrahim, J. G. and Molenberghs, G. Missing data methods in longitudinal studies: A review. *Test*, 18(1):1–43, 2009.

Ida, Y., Fujiwara, Y., and Kashima, H. Fast sparse group lasso. *Advances in Neural Information Processing Systems*, 32, 2019.

Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 433–440, 2009.

Kale, S., Karnin, Z., Liang, T., and Pál, D. Adaptive feature selection: Computationally efficient online sparse linear regression under rip. In *International Conference on Machine Learning*, pp. 1780–1788, 2017.

Kukliansky, D. and Shamir, O. Attribute efficient linear regression with distribution-dependent sampling. In *International Conference on Machine Learning*, pp. 153–161, 2015.

Madani, O., Lizotte, D. J., and Greiner, R. Active model selection. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 357–365, 2004.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT press, 2018.

Price, B. S. and Sherwood, B. A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, 18(232):1–39, 2018.

Rostamizadeh, A., Agarwal, A., and Bartlett, P. Learning with missing features. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 635–642, Arlington, Virginia, USA, 2011. AUAI Press.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Su, Y., Gao, X., Li, X., and Tao, D. Multivariate multilinear regression. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(6):1560–1573, 2012.

Tan, Q., Yu, G., Domeniconi, C., Wang, J., and Zhang, Z. Incomplete multi-view weak-label learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2703–2709, 2018.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

Watt, J., Borhani, R., and Katsaggelos, A. K. *Machine Learning Refined: Foundations, Algorithms, and Applications*. 2nd edition, Cambridge University Press, 2020.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Zhu, H., Li, T., and Zhao, B. Statistical learning methods for neuroimaging data analysis with applications. *Annual Review of Biomedical Data Science*, 6:73–104, 2023.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.

# Supplementary Material

In Section A.1, we present the pseudo code for the multivariate AERR algorithm mentioned in Section 5.1. Section A.2 provides additional experiment results for Algorithm 4. Section B shows another real data application of the proposed method. The proof for all theorems and lemmas are given in Section C.

## A. Additional Experiment Results

### A.1. The multivariate AERR

We use $\boldsymbol{C}_{t,:j}$ to denote the $j$-th column vector of matrix $\boldsymbol{C}_t$. Let $C_{t,j':j}$ denote the $(j', j)$ element of $\boldsymbol{C}_t$. The following algorithm extends AERR (Hazan & Koren, 2012) to multivariate settings, which is identical to AERR when $q = q_0 = 1$.

The number of the observed features in this algorithm may differ from Algorithm 1. On one hand, $\tilde{\boldsymbol{x}}_t$ in Line 3 is constructed by sampling with replacement, which differs from Line 6 in Algorithm 1. Thus, we observe at most $p_0 - 1$ attributes in this step, whereas we observe exactly $p_0 - 1$ attributes in that step of Algorithm 1. On the other hand, due to the "for loop" in Lines 7-11 we may need to observe at most $q_0$ additional attributes; however, due to Line 7 of Algorithm 1 we may only need to observe at most one extra attribute. When $q_0$ is small, as in the setting of Section 5.1, the total number of observed feature attributes may be smaller than in Algorithm 1.

**Algorithm 5** Multivariate AERR

1: **Initialize:** $B, \eta > 0$, $\boldsymbol{W}_1$ satisfying $0 < \|\boldsymbol{W}_1\|_F \leq B$, $q_0 \geq 1$, $p_0 \geq 2$
2: **for** $t = 1, \ldots, T$ **do**
3: Choose $\{j_{t,r} : r \in [p_0 - 1]\}$ uniformly from $[p]$ with replacement
4: $\tilde{\boldsymbol{x}}_t \leftarrow \frac{p}{p_0 - 1} \sum_{r=1}^{p_0-1} x_{t,j_{t,r}} \boldsymbol{e}_{j_{t,r}}^{[p]}$
5: $\boldsymbol{V}_t \leftarrow \boldsymbol{W}_t$
6: Choose $\{j_{t,r} : r \in [q_0]\}$ uniformly from $[q]$ without replacement
7: **for** $r = 1, \ldots, q_0$ **do**
8: Pick an index $j_t'$ with probability $\frac{(W_{t,j_t':j_{t,r}})^2}{\|\boldsymbol{W}_{t,:j_{t,r}}\|^2}$ from $[p]$ and observe $x_{t,j_t'}$
9: $\breve{\boldsymbol{g}}_t \leftarrow \tilde{\boldsymbol{x}}_t(\|\boldsymbol{W}_{t,:j_{t,r}}\|^2 x_{t,j_t'} / W_{t,j_t':j_{t,r}} - y_{t,j_{t,r}})$
10: $\boldsymbol{V}_{t,:j_{t,r}} \leftarrow \boldsymbol{W}_{t,:j_{t,r}} - \eta \breve{\boldsymbol{g}}_t$
11: **end for**
12: $\boldsymbol{W}_{t+1} \leftarrow B\boldsymbol{V}_t / \max\{\|\boldsymbol{V}_t\|_F, B\}$
13: **end for**
14: **Return:** $\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{W}_t$

### A.2. Performance of Algorithm 4

We present additional experiment results for Algorithm 4 here. Let $q = 100$, $p = 5$, and $p_0 = 2$. Set all elements of the last 90 rows of the true coefficient matrix $\boldsymbol{W}_0$ to 0 and the remaining elements are uniformly sampled from $\{0, 1, -1, 2, -2\}$. The remaining data generation procedure is similar to that in Section 5.2, where $\sigma_\varepsilon^2 = 2$. For each of the 300 replicates, we generate $T' = 5000$ new samples without outliers, denoted as $\{\boldsymbol{y}_i^*, \boldsymbol{x}_i^*\}_{i=1}^{T'}$, and compute the mean prediction error (abbreviated as PE), defined as $\frac{1}{T'} \sum_{i=1}^{T'} \|\boldsymbol{y}_i^* - \hat{\boldsymbol{W}}^\top \boldsymbol{x}_i^*\|_1$, and the mean prediction error with penalty (abbreviated as PEP), defined as $\frac{1}{T'} \sum_{i=1}^{T'} \|\boldsymbol{y}_i^* - \hat{\boldsymbol{W}}^\top \boldsymbol{x}_i^*\|_1 + P_\lambda(\hat{\boldsymbol{W}}_2)$, where $\hat{\boldsymbol{W}}$ is the output of Algorithm 4. For comparison, we also conduct Algorithms 3 and 2 and calculate the corresponding mean prediction error (abbreviated as PE0 and PE*, respectively). We set $B = 100$ for all methods and let $\lambda_1 = 0.1$ and $\lambda_2 = 0.001$ for Algorithms 2 and 4.

Table S1 presents the means (and standard variations) of PEP, PE, PE0, and PE* values over 300 replicates under various sample sizes of $T$ and observable numbers of features $p_0$. Both PEP and PE show slight reduction as $T$ increases. As indicated by Theorem 4.2, we cannot expect the prediction error to be arbitrarily precise as $T \to \infty$.

The comparison between PE and PE0 serves as a supplement of Section 5.3. It is observed that when $p_0$ is relatively small, penalization may be beneficial for achieving more accurate predictions, especially when $T$ is not large. However, if $p_0$ is large enough, the performance of Algorithms 3 and 4 is comparable even when $T$ is not large. The comparison between

*Table S1.* Experiment results: Means (and standard variations) of PEP, PE, PE0 and PE* over 300 replicates.

| $p_0$ | | $T = 1 \times 10^4$ | $T = 2 \times 10^4$ | $T = 5 \times 10^4$ |
|---|---|---|---|---|
| 30 | PEP | 12.25 (0.61) | 11.85 (0.56) | 11.57 (0.61) |
| | PE | 9.80 (0.66) | 9.67 (0.60) | 9.63 (0.68) |
| | PE0 | 12.25 (1.15) | 11.48 (1.21) | 11.04 (1.24) |
| | PE* | 42.82 (1.75) | 41.37 (1.63) | 39.30 (1.68) |
| 60 | PEP | 11.03 (0.41) | 10.51 (0.38) | 10.20 (0.39) |
| | PE | 8.11 (0.38) | 7.84 (0.33) | 7.71 (0.35) |
| | PE0 | 8.41 (0.36) | 7.76 (0.34) | 7.45 (0.34) |
| | PE* | 44.20 (1.91) | 42.42 (1.89) | 39.42 (1.60) |

PE and PE* supplements Section 5.2, clearly demonstrating that Algorithm 4 outperforms Algorithm 2 in the presence of outliers.

## B. Another Real Data Application: Children Activity Data

To prevent childhood obesity, researchers investigated the relationship between children's daily physical activity and potential risk factors. The dataset can be accessed at `http://jeffgoldsmith.com/IWAFDA/shortcourse_data.html`. The study involved 420 participants recruited from various Head Start centers. These participants wore accelerometers to monitor their body activity intensity, resulting in 144 daily observations of outcomes for each child. The dataset includes 15 features for each participant, including BMI Z-score, three types of skinfold thicknesses, age at recruitment, sex, season of the study, presence of asthma diagnosis, mother's birthplace and education, work status, number of rooms at home, and two behavioral variables related to daily time spent on TV and video games (Rundle et al., 2009; Lovasi et al., 2011).

We standardized the data and set the constraint parameter $B$ to 5. The other settings remain consistent with those described in Section 5.4. The mean prediction error was computed for different values of $p_0 \in \{3, 6, 9, 12\}$ and $q_0 \in \{20, 50, 80\}$, and the results are presented in Figure S1. The observed trends closely resemble those depicted in Figure 1 in the main text, confirming the superiority of our proposed method over the multivariate AERR.

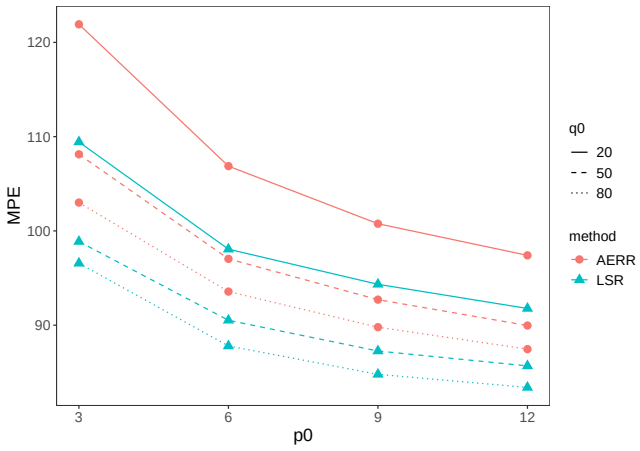


*Figure S1.* Application results for children activity data: The mean prediction error for $p_0 = 3, 6, 9, 12$, and $q_0 = 20$ (solid line), 50 (dash line), 80 (dotted line), using Algorithm 1 (LSR, blue triangle) or the "multivariate AERR" (red dot).

# C. Technical Proofs

## C.1. Proofs of Theorems 3.1 and 3.3

To prove Theorems 3.1 and 3.3, we acknowledge two types of randomness involved in the output $\hat{\boldsymbol{W}}$ of an algorithm. The first arises from the algorithm, denoted as $A$, and the second is due to dependence on the associated random sample $S \triangleq \{(\boldsymbol{x}_t, \boldsymbol{y}_t) : t = 1, 2, \ldots, T\}$. Give the algorithm, we use $\mathbb{E}_A(\cdot|S)$ to represent the expectation of the associated quantity taken with respect to $A$ conditional on the sample data $S$. Similarly, for $t = 1, \ldots, T$, we let $A_t$ represent the randomness arising from the $t$-th iteration of the algorithm, let $S_t$ denote the $t$-th random sample $(\boldsymbol{x}_t, \boldsymbol{y}_t)$, and let $\mathbb{E}_{A_t}(\cdot|S_t)$ represent the conditional expectation with respect to $A_t$ given $S_t$. As $A$ and $S$ encompass all random objects under consideration but $A_t$ and $S_t$ do not, $\mathbb{E}_A(\cdot|S)$ is essentially $\mathbb{E}(\cdot|S)$, whereas $\mathbb{E}_{A_t}(\cdot|S_t)$ is not always $\mathbb{E}(\cdot|S_t)$.

We first present the following lemmas whose proofs are given in Section C.3.

**Lemma C.1.** *Let $f : \mathbb{R}^{p\times q} \to \mathbb{R}$ be a convex function. Then the following results hold:*

*(a). for any $p \times q$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,*

$$f(\boldsymbol{A}) - f(\boldsymbol{B}) \le tr\{(\boldsymbol{A} - \boldsymbol{B})^\mathsf{T}\boldsymbol{D}\}, \quad \text{for any } \boldsymbol{D} \in \partial f(\boldsymbol{A}),$$

*where*

$$\partial f(\boldsymbol{A}) \triangleq \{\boldsymbol{D} \in \mathbb{R}^{p\times q} : f(\boldsymbol{Z}) \ge f(\boldsymbol{A}) + tr\{(\boldsymbol{Z} - \boldsymbol{A})^\mathsf{T}\boldsymbol{D}\} \text{ for all } \boldsymbol{Z}\}$$

*is the subdifferential of $f$ at $\boldsymbol{A}$.*

*(b). if $f$ is differentiable at $\boldsymbol{A}$, then*

$$\partial f(\boldsymbol{A}) = \{\nabla f(\boldsymbol{A})\},$$

*where $\nabla f(\boldsymbol{A})$ is the gradient of $f$ at $\boldsymbol{A}$, defined as the $p \times q$ matrix with the $(j,k)$ element given by $\frac{\partial f(\boldsymbol{A})}{\partial A_{jk}}$ for $1 \le j \le p$ and $1 \le k \le q$. Here $A_{jk}$ represents the $(j,k)$ element of $\boldsymbol{A}$.*

**Lemma C.2.** *Consider Algorithm 1 and the given sample $S$. Assume the conditions of Theorem 3.1 hold. For any $\|\boldsymbol{W}^*\|_F \le B$, we have that*

$$\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T} tr\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)^\mathsf{T}\}\middle| S\right] \le \frac{2B^2}{\eta T} + \frac{\eta B_x^2 p}{p_0 - 1}\left(\frac{B_y^2 q}{q_0} + B^2 B_x^2\right).$$

**Lemma C.3.** *Consider Algorithm 2 and the given sample $S$. Assume the conditions of Theorem 3.3 hold. For any $\|\boldsymbol{W}^*\|_F \le B$, we have that*

$$\begin{aligned}&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[tr\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)\right\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right)\\ \le &\frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)B_x^2 \cdot \frac{p}{p_0 - 1}\left(B_y^2 \cdot \frac{q}{q_0} + B^2 B_x^2\right) + \frac{1}{T}P_\lambda(\boldsymbol{W}_1).\end{aligned}$$

*Proof of Theorem 3.1.* By the definition of $\boldsymbol{W}_{t+1}$ in Algorithm 1, it is immediate that $\|\boldsymbol{W}_{t+1}\|_F \le B$, and thus, $\|\hat{\boldsymbol{W}}\|_F \le B$ by the triangle inequality.

(a). For $t \geq 1$ and $(\boldsymbol{x}_t, \boldsymbol{y}_t) \in S$, define $L_t(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W}^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t\|^2$. Then

$$\begin{aligned}\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} L_t(\boldsymbol{W}_t)\right\} &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{L_t(\boldsymbol{W}_t)\} \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}\{L_t(\boldsymbol{W}_t)|\boldsymbol{W}_t\}] \\ &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{R(\boldsymbol{W}_t)\} \\ &= \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} R(\boldsymbol{W}_t)\right\}, \end{aligned} \tag{11}$$

where the second step is due to the law of iterative expectations, and the third step is due to $\mathbb{E}\{L_t(\boldsymbol{W}_t)|\boldsymbol{W}_t\} = R(\boldsymbol{W}_t)$, since by the construction in Algorithm 1, $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ in $L_t(\boldsymbol{W}_t)$ are independent of $\boldsymbol{W}_t$. Here, the expectations are evaluated with respect to the joint distributions for the associated random variables.

For any given $(\boldsymbol{x}_t, \boldsymbol{y}_t) \in S$ and $\boldsymbol{W}^*$ with $\|\boldsymbol{W}^*\|_F \leq B$, the convexity of $L_t(\boldsymbol{W})$ with respect to $\boldsymbol{W}$ leads to

$$L_t(\boldsymbol{W}_t) - L_t(\boldsymbol{W}^*) \leq \mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)^\mathsf{T}\} \tag{12}$$

by Lemma C.1, where $\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)^\mathsf{T}$ is in fact the gradient of $L_t(\boldsymbol{W})$ at $\boldsymbol{W}_t$. Averaging (12) for $t = 1, \ldots, T$ and then taking the conditional expectation given $A$ and $S$, we have that

$$\begin{aligned}\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\{L_t(\boldsymbol{W}_t) - L_t(\boldsymbol{W}^*)\}\middle| S\right] &\leq \mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)^\mathsf{T}\}\middle| S\right] \\ &\leq \frac{2B^2}{\eta T} + \eta B_x^2 \cdot \frac{p}{p_0 - 1}\left(B_y^2 \cdot \frac{q}{q_0} + B^2 B_x^2\right), \end{aligned} \tag{13}$$

where the second step follows from Lemma C.2.

Noting that by definition in Section 3.1, for any $1 \leq t \leq T$, we have that

$$\mathbb{E}\{L_t(\boldsymbol{W}^*)\} = R(\boldsymbol{W}^*). \tag{14}$$

Consequently, by taking expectation of (13) with respect to $S$ and using (14), we arrive at

$$\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} L_t(\boldsymbol{W}_t)\right\} \leq R(\boldsymbol{W}^*) + \frac{2B^2}{\eta T} + \frac{p\eta B_x^2}{p_0 - 1}\left(\frac{qB_y^2}{q_0} + B^2 B_x^2\right). \tag{15}$$

Note that

$$\mathbb{E}\{R(\hat{\boldsymbol{W}})\} = \mathbb{E}\left\{R\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{W}_t\right)\right\} \leq \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} R(\boldsymbol{W}_t)\right\} = \mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} L_t(\boldsymbol{W}_t)\right\}, \tag{16}$$

where the second step is due to the convexity of $R(\boldsymbol{W})$ with respect to $\boldsymbol{W}$, and the last step is due to (11). Combining this inequality with (15) proves result (b).

(b). By the choice of $\eta$ in (5) of Theorem 3.1(c),

$$\begin{aligned}\frac{2B^2}{\eta T} + \eta B_x^2 \cdot \frac{p}{p_0 - 1}\left(B_y^2 \cdot \frac{q}{q_0} + B^2 B_x^2\right) &= 2\sqrt{2}BB_x\sqrt{\frac{1}{T}\cdot\frac{p}{p_0 - 1}\cdot\left(B_y^2 \cdot \frac{q}{q_0} + B^2 B_x^2\right)} \\ &\leq 2\sqrt{2}\tilde{B}^2\sqrt{\frac{1}{T}\cdot\frac{p}{p_0 - 1}\cdot\left(\frac{q}{q_0} + 1\right)} \\ &\leq 4\tilde{B}^2\sqrt{\frac{1}{T}\cdot\frac{p}{p_0 - 1}\cdot\frac{q}{q_0}}, \end{aligned} \tag{17}$$

where the last step is due to $q/q_0 \geq 1$. Therefore, (15) can be rewritten as

$$\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T} L_t(\boldsymbol{W}_t)\right\} \leq R(\boldsymbol{W}^*) + 4\tilde{B}^2\sqrt{\frac{1}{T}\cdot\frac{p}{p_0-1}\cdot\frac{q}{q_0}},$$

which proves (c) using the result of (b). □

*Proof of Theorem 3.3.* The bound of the output can be proved identical to the proof of Theorem 3.1(a).

(a). Modifying the derivation for (13) by adding the penalty function, we obtain that by (12) and Lemma C.3,

$$\begin{aligned}
&\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\{L_t(\boldsymbol{W}_t) + P_\lambda(\boldsymbol{W}_t) - L_t(\boldsymbol{W}^*) - P_\lambda(\boldsymbol{W}^*)\}\middle| S\right] \\
\leq &\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t)\right\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right) \\
\leq &\frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)B_x^2\cdot\frac{p}{p_0-1}\left(B_y^2\cdot\frac{q}{q_0} + B^2B_x^2\right) + \frac{1}{T}P_\lambda(\boldsymbol{W}_1).
\end{aligned} \tag{18}$$

Taking the expectation of (18) with respect to $S$ and using (14) yields that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{L_t(\boldsymbol{W}_t) + P_\lambda(\boldsymbol{W}_t)\}\right] \leq R(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + \Delta_1,$$

where the expectation is taken with respect to all sources of randomness, involving both $A$ and $S$. Analogous to (16) and by the convexity of $P_\lambda(\cdot)$, we have that

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{L_t(\boldsymbol{W}_t) + P_\lambda(\boldsymbol{W}_t)\}\right].$$

Combining the above two inequalities proves result (b).

(b). Suppose that $\eta_t$ is set as the constant $\eta$ in (5) for each $t = 1, \ldots, T$. Then using the same argument as (17), $\Delta_1$ can be bounded by

$$4\tilde{B}^2\sqrt{\frac{1}{T}\cdot\frac{p}{p_0-1}\cdot\frac{q}{q_0}} + \frac{1}{T}P_\lambda(\boldsymbol{W}_1),$$

and result (c) is then immediate due to Theorem 3.3(b).

(c). Define $\boldsymbol{W}_0^*$ to satisfy

$$R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*) = \inf_{\|\boldsymbol{W}\|_F \leq B}\{R(\boldsymbol{W}) + P_\lambda(\boldsymbol{W})\}. \tag{19}$$

If $\eta_T T \to \infty$ and $\frac{1}{T}\sum_{t=1}^{T}\eta_t \to 0$ as $T \to \infty$, the $\Delta_1 \to 0$ as $T \to \infty$. Consequently, by Theorem 3.3(b), we have that

$$\lim_{T\to\infty}\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \leq R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*).$$

By definition of $\boldsymbol{W}_0^*$ in (19) and Theorem 3.3(a), we have that given $T$, $S$ and $A$,

$$R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}}) \geq R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*)$$

for any output $\hat{\boldsymbol{W}}$, yielding that

$$\mathbb{E}\{R(\hat{\boldsymbol{W}}) + P_\lambda(\hat{\boldsymbol{W}})\} \geq R(\boldsymbol{W}_0^*) + P_\lambda(\boldsymbol{W}_0^*)$$

for any $T$. Combining the proceeding inequalities proves the result. □

### C.2. Proof of Theorems 4.1 and 4.2

Consider Algorithm 3 or 4 and the given sample $S$. For any matrix $\boldsymbol{W}$, define

$$L_t(\boldsymbol{W}) = \|\boldsymbol{W}^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t\|_1 \text{ and } \tilde{L}_t(\boldsymbol{W}) = \frac{q}{q_0}\mathbb{E}_{A_t}(\|\boldsymbol{W}_{:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1|S_t), \tag{20}$$

where the use of $A_t$ emphasizes the algorithm randomness at step $t$ and $S_t = (\boldsymbol{x}_t, \boldsymbol{y}_t)$. To prove Theorem 4.1, we first introduce two lemmas whose proofs are given in C.3.

**Lemma C.4.** *Consider Algorithm 3 and the given sample $S$. Assume the conditions of Theorem 4.1 hold. For any $\|\boldsymbol{W}^*\|_F \le B$, we have that*

$$\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T} tr\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\}\middle| S\right] \le \frac{2B^2}{\eta T} + \frac{q^2 p_0}{2pq_0}B_x^2\eta.$$

**Lemma C.5.** *Assume the conditions of Theorem 4.1 or Theorem 4.2 hold. Then for any $t \ge 1$ and any matrix $\boldsymbol{W}$ satisfying $\|\boldsymbol{W}\|_F \le B$, we have that*

$$|\tilde{L}_t(\boldsymbol{W}) - L_t(\boldsymbol{W})| \le \sqrt{q}\, BB_x\sqrt{1 - \frac{p_0}{p}},$$

*where $\tilde{L}_t(\cdot)$ and $L_t(\cdot)$ are given by (20).*

**Lemma C.6.** *Consider Algorithm 4 and the given sample $S$. Assume the conditions of Theorem 4.2 hold. For any $\|\boldsymbol{W}^*\|_F \le B$, we have that*

$$\begin{aligned}&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[tr\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right)\\ \le &\frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2 p_0}{2pq_0}B_x^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1).\end{aligned}$$

*Proof of Theorem 4.1.* The bound of the output can be verified identical to the proof of Theorem 3.1.

(a). We first show that $\mathbb{E}_{A_t}(\breve{\boldsymbol{G}}_t|S_t) \in \partial\tilde{L}_t(\boldsymbol{W}_t)$, where $\partial\tilde{L}_t(\boldsymbol{W}_t)$ represents the subdifferential of $\tilde{L}_t(\cdot)$ in (20) at $\boldsymbol{W}_t$. Indeed, $\tilde{L}_t(\boldsymbol{W}_t)$ can be explicitly written as

$$\tilde{L}_t(\boldsymbol{W}_t) = \frac{q}{q_0}\mathbb{E}_{A_t}(\|\boldsymbol{W}_{t,:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1|S_t),$$

since $\boldsymbol{W}_t$ is independent of $A_t$ or $S_t$. Therefore, for any given $\boldsymbol{W}$,

$$\begin{aligned}\tilde{L}_t(\boldsymbol{W}) - \tilde{L}_t(\boldsymbol{W}_t) &= \frac{q}{q_0}\mathbb{E}_{A_t}\left(\|\boldsymbol{W}_{:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1 - \|\boldsymbol{W}_{t,:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1|S_t\right)\\ &\ge \frac{q}{q_0}\mathbb{E}_{A_t}\left[\mathrm{tr}\{(\boldsymbol{W}_{:\mathcal{O}_t} - \boldsymbol{W}_{t:\mathcal{O}_t})^\mathsf{T}\tilde{\boldsymbol{x}}_t\boldsymbol{\phi}_{t,\mathcal{O}_t}^\mathsf{T}\}|S_t\right]\\ &= \frac{q}{q_0}\mathbb{E}_{A_t}\left[\mathrm{tr}\{(\boldsymbol{W} - \boldsymbol{W}_t)^\mathsf{T}\tilde{\boldsymbol{x}}_t\boldsymbol{\phi}_t^\mathsf{T}\}|S_t\right]\\ &= \mathrm{tr}\{(\boldsymbol{W} - \boldsymbol{W}_t)^\mathsf{T}\mathbb{E}_{A_t}(\breve{\boldsymbol{G}}_t|S_t)\},\end{aligned}$$

where the second step is due to Lemma C.1, the convexity of $\|\cdot\|_1$, the construction of $\boldsymbol{\phi}_{t,\mathcal{O}_t}$ in Algorithm 3, and the calculation of subgradient $\|\boldsymbol{W}_{:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1$ at $\boldsymbol{W}_{t,:\mathcal{O}_t}$; the third step comes from the fact that the components of $\boldsymbol{\phi}_t$ on the complement of $\mathcal{O}_t$ is zero; and the last step uses the construction of $\breve{\boldsymbol{G}}_t$ in Algorithm 3. This proves that $\mathbb{E}_{A_t}(\breve{\boldsymbol{G}}_t|S_t) \in \partial\tilde{L}_t(\boldsymbol{W}_t)$ by definition as stated in Lemma C.1(a).

Note that $\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S) = \mathbb{E}_{A_t}(\breve{\boldsymbol{G}}_t|S_t)$, since, by construction in Algorithm 3, the randomness of $\breve{\boldsymbol{G}}_t$ solely comes from $A_t$, $\boldsymbol{W}_t$ and $S_t$, and thus, $\breve{\boldsymbol{G}}_t$ is independent of $S\backslash S_t$ or $A\backslash A_t$, conditionally on $\boldsymbol{W}_t$ and $S_t$. Hence, $\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S) \in \partial\tilde{L}_t(\boldsymbol{W}_t)$, giving that

$$\tilde{L}_t(\boldsymbol{W}_t) - \tilde{L}_t(\boldsymbol{W}^*) \le \mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S)\}. \tag{21}$$

Averaging (21) for $t = 1, \ldots, T$ and then taking the expectation with respect to $A$ conditionally on $S$ yields that

$$\begin{aligned}\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\left\{\tilde{L}_t(\boldsymbol{W}_t) - \tilde{L}_t(\boldsymbol{W}^*)\right\}\middle| S\right] &\le \mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S)\}\middle| S\right] \\ &= \mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\}\middle| S\right] \\ &\le \frac{2B^2}{\eta T} + \frac{q^2 p_0}{2pq_0}B_x^2\eta,\end{aligned}$$

where the second step is due to the law of iterative expectations, and the third step is due to Lemma C.4. As a consequence of Lemma C.5, we further have that

$$\mathbb{E}_A\left\{\frac{1}{T}\sum_{t=1}^{T}L_t(\boldsymbol{W}_t)\middle| S\right\} \le \frac{1}{T}\sum_{t=1}^{T}L_t(\boldsymbol{W}^*) + \frac{2B^2}{\eta T} + \frac{q^2 p_0}{2pq_0}B_x^2\eta + 2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right),$$

which, by taking expectation with respect to $S$ and by (11), gives that

$$\mathbb{E}\left\{\frac{1}{T}\sum_{t=1}^{T}R(\boldsymbol{W}_t)\right\} \le R(\boldsymbol{W}^*) + \frac{2B^2}{\eta T} + \frac{q^2 p_0}{2pq_0}B_x^2\eta + 2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right).$$

Result (b) is then proved due to the convexity of $R(\boldsymbol{W})$ with respect to $\boldsymbol{W}$.

(b). Set $\eta = \frac{2B}{B_x q}\sqrt{\frac{pq_0}{p_0 T}}$ as required in Theorem 4.1. The proof follows the same steps as in (b) with all terms $\frac{2B^2}{\eta T} + \frac{q^2 p_0}{2pq_0}B_x^2\eta$ replaced by $2BB_x\sqrt{\frac{q^2 p_0}{pq_0 T}}$. □

*Proof of Theorem 4.2.* The proof of the bound is identical to the proof of Theorem 3.1.

(a). The proof of Theorem 4.2(b) combines the proofs of Theorems 3.3 and 4.1. Indeed, by the law of iterative expectations and Lemma C.6,

$$\begin{aligned}&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\mathbb{E}(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S)\right\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right) \\ =&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\right\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right) \\ \le&\frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2 p_0}{2pq_0}B_x^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1).\end{aligned}$$

Then by (21), we have that

$$\begin{aligned}&\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\left\{\tilde{L}_t(\boldsymbol{W}_t) - \tilde{L}_t(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right\}\middle| S\right] \\ \le&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t, S)\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right]\middle| S\right) \\ \le&\frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2 p_0}{2pq_0}B_x^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1),\end{aligned}$$

which, by Lemma C.5, gives that

$$\begin{aligned}\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T}\{L_t(\boldsymbol{W}_t) + P_\lambda(\boldsymbol{W}_t)\}\middle| S\right] \le& \frac{1}{T}\sum_{t=1}^{T}L_t(\boldsymbol{W}^*) + P_\lambda(\boldsymbol{W}^*) + 2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right) \\ &+ \frac{2B^2}{\eta_T T} + \left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2 p_0}{2pq_0}B_x^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1).\end{aligned}$$

Taking expectation with respect to $S$, we arrive at

$$\mathbb{E}\left\{R(\hat{\boldsymbol{W}})+P_\lambda(\hat{\boldsymbol{W}})\right\} \leq R(\boldsymbol{W}^*)+P_\lambda(\boldsymbol{W}^*)+2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right) + \frac{2B^2}{\eta_T T}+\left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2p_0}{2pq_0}B_x^2+\frac{1}{T}P_\lambda(\boldsymbol{W}_1), \tag{22}$$

due to the convexity of $R(\boldsymbol{W})+P_\lambda(\boldsymbol{W})$ with respect to $\boldsymbol{W}$.

(b). Assume that $\eta_t=\eta=\frac{2B}{B_x q}\sqrt{\frac{pq_0}{p_0 T}}$ for all $t=1,\ldots,T$, the last line of (22) can be rewritten as $2BB_x\sqrt{\frac{q^2p_0}{pq_0T}}+\frac{1}{T}P_\lambda(\boldsymbol{W}_1)$.

(c). Assume that $\eta_T T\to\infty$ and $\frac{1}{T}\sum_{t=1}^{T}\eta_t\to 0$ as $T\to\infty$, the last line of (22) converges to 0 as $T\to\infty$. Consequently,

$$\limsup_{T\to\infty}\mathbb{E}\{R(\hat{\boldsymbol{W}})+P_\lambda(\hat{\boldsymbol{W}})\}\leq R(\boldsymbol{W}^*)+P_\lambda(\boldsymbol{W}^*)+2\sqrt{q}BB_x\left(\sqrt{1-\frac{p_0}{p}}\right).$$

□

### C.3. Proofs of Lemmas C.1-C.6 and Proposition 3.2

*Proof of Lemma C.1.* By the definition of $\partial f(\boldsymbol{A})$, the result (a) for the general case follows immediately. Now consider the special case (b) in which $f$ is differentiable at $\boldsymbol{A}$. The goal is to prove $\partial f(\boldsymbol{A})=\{\nabla f(\boldsymbol{A})\}$.

Let $\tilde{\boldsymbol{A}}=\mathrm{vec}(\boldsymbol{A})\in\mathbb{R}^{pq}$ denote the vectorization of matrix $\boldsymbol{A}$. Define $g(\tilde{\boldsymbol{A}})=f(\boldsymbol{A})$. It is immediate that $g$ is also convex over the vector space $\mathbb{R}^{pq}$ and that $g$ is differentiable at $\tilde{\boldsymbol{A}}$. The subdifferential of $g$ at $\tilde{\boldsymbol{A}}$ (Mohri et al. (2018, Definition B.31); Beck (2017, Definition 3.2)) is given by

$$\partial g(\tilde{\boldsymbol{A}})\triangleq\{\tilde{\boldsymbol{D}}\in\mathbb{R}^{pq}: f(\tilde{\boldsymbol{Z}})\geq f(\tilde{\boldsymbol{A}})+(\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{A}})^{\mathsf{T}}\tilde{\boldsymbol{D}}\text{ for all }\tilde{\boldsymbol{Z}}\}.$$

Simple algebra yields that the vectorization characterizes a one-to-one mapping between $\partial g(\tilde{\boldsymbol{A}})$ and $\partial f(\boldsymbol{A})$. By Mohri et al. (2018, Lemma B.32), $\partial g(\tilde{\boldsymbol{A}})=\{\nabla g(\tilde{\boldsymbol{A}})\}$, where $\nabla g(\tilde{\boldsymbol{A}})$ represents the gradient of $g$ evaluated at $\tilde{\boldsymbol{A}}$. Furthermore, the vectorization of $\nabla f(\boldsymbol{A})$ is $\nabla g(\tilde{\boldsymbol{A}})$. Hence, we conclude that $\partial f(\boldsymbol{A})=\{\nabla f(\boldsymbol{A})\}$. □

*Proof of Lemma C.2.* By the condition $\|\boldsymbol{W}^*\|_F\leq B$ and the definitions of $\boldsymbol{V}_t$ and $\boldsymbol{W}_{t+1}$ in Algorithm 1, we have that

$$\begin{aligned}\|\boldsymbol{W}_{t+1}-\boldsymbol{W}^*\|_F^2 &\leq \|\boldsymbol{V}_t-\boldsymbol{W}^*\|_F^2\\ &=\|\boldsymbol{W}_t-\eta\breve{\boldsymbol{G}}_t-\boldsymbol{W}^*\|_F^2\\ &=\|\boldsymbol{W}_t-\boldsymbol{W}^*\|_F^2-2\eta\cdot\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^{\mathsf{T}}\breve{\boldsymbol{G}}_t\}+\eta^2\|\breve{\boldsymbol{G}}_t\|_F^2,\end{aligned}$$

yielding that

$$\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^{\mathsf{T}}\breve{\boldsymbol{G}}_t\}\leq\frac{1}{2\eta}\left\{\|\boldsymbol{W}_t-\boldsymbol{W}^*\|_F^2-\|\boldsymbol{W}_{t+1}-\boldsymbol{W}^*\|_F^2\right\}+\frac{\eta}{2}\|\breve{\boldsymbol{G}}_t\|_F^2.$$

Thus,

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^{\mathsf{T}}\breve{\boldsymbol{G}}_t\}\leq\frac{2B^2}{\eta T}+\frac{\eta}{2T}\sum_{t=1}^{T}\|\breve{\boldsymbol{G}}_t\|_F^2, \tag{23}$$

since $0\leq\|\boldsymbol{W}_t-\boldsymbol{W}^*\|_F^2\leq 4B^2$ for any $t$.

By the construction of $\tilde{\boldsymbol{x}}_t$, it is direct to prove that each coordinate of $\tilde{\boldsymbol{x}}_t$ is an unbiased estimate of the corresponding coordinate of $\boldsymbol{x}_t$, and thus, $\mathbb{E}_A(\tilde{\boldsymbol{x}}_t|S)=\boldsymbol{x}_t$. Further, we can also verify that $\mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t\|^2|S)=\frac{p}{p_0-1}\|\boldsymbol{x}_t\|^2$. Analogously, $\mathbb{E}_A(\tilde{\boldsymbol{y}}_t|S)=\boldsymbol{y}_t$ and $\mathbb{E}_A(\|\tilde{\boldsymbol{y}}_t\|^2|S)=\frac{q}{q_0}\|\boldsymbol{y}_t\|^2$. Hence, according to the definition of $\breve{\boldsymbol{G}}_t$ and $x_{t,j_t}$ in Algorithm 1, we obtain

that

$$\begin{aligned}\mathbb{E}_A(\breve{\boldsymbol{G}}_t|\boldsymbol{W}_t,S) &= \mathbb{E}_A(\tilde{\boldsymbol{x}}_t|\boldsymbol{W}_t,S)\mathbb{E}_A\{(x_{t,j_t}\|\boldsymbol{W}_t\|_F^2\boldsymbol{W}_{t,j_t:}^\mathsf{T}/\|\boldsymbol{W}_{t,j_t:}^\mathsf{T}\|^2-\tilde{\boldsymbol{y}}_t)^\mathsf{T}|\boldsymbol{W}_t,S\}\\ &= \boldsymbol{x}_t\left(\sum_{k=1}^p x_{t,k}\boldsymbol{W}_{t,k:}^\mathsf{T}-\boldsymbol{y}_t\right)^\mathsf{T}\\ &= \boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t-\boldsymbol{y}_t)^\mathsf{T},\end{aligned} \tag{24}$$

where the second step is due to the independence between $(\tilde{\boldsymbol{x}}_t,\tilde{\boldsymbol{y}}_t)$ and $\boldsymbol{W}_t$. Therefore, by the law of iterative expectations,

$$\begin{aligned}\mathbb{E}_A[\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\}|S] &= \mathbb{E}_A(\mathbb{E}_A[\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\}|\boldsymbol{W}_t,S]|S)\\ &= \mathbb{E}_A[\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t-\boldsymbol{y}_t)\}|S].\end{aligned} \tag{25}$$

In addition, by the identity

$$\begin{aligned}\|\breve{\boldsymbol{G}}_t\|_F^2 &= \|\tilde{\boldsymbol{x}}_t(x_{t,j_t}\|\boldsymbol{W}_t\|_F^2\boldsymbol{W}_{t,j_t:}^\mathsf{T}/\|\boldsymbol{W}_{t,j_t:}^\mathsf{T}\|^2-\tilde{\boldsymbol{y}}_t)^\mathsf{T}\|_F^2\\ &= \|\tilde{\boldsymbol{x}}_t\|^2\left\|x_{t,j_t}\|\boldsymbol{W}_t\|_F^2\boldsymbol{W}_{t,j_t:}^\mathsf{T}/\|\boldsymbol{W}_{t,j_t:}^\mathsf{T}\|^2-\tilde{\boldsymbol{y}}_t\right\|^2,\end{aligned}$$

we obtain that

$$\begin{aligned}\mathbb{E}_A(\|\breve{\boldsymbol{G}}_t\|_F^2|S) &= \mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t\|^2|S)\mathbb{E}_A\left(\left\|x_{t,j_t}\|\boldsymbol{W}_t\|_F^2\boldsymbol{W}_{t,j_t:}^\mathsf{T}/\|\boldsymbol{W}_{t,j_t:}^\mathsf{T}\|^2-\tilde{\boldsymbol{y}}_t\right\|^2|S\right)\\ &\le 2\|\boldsymbol{x}_t\|^2\cdot\frac{p}{p_0-1}\left\{\|\boldsymbol{y}_t\|^2\cdot\frac{q}{q_0}+\mathbb{E}_A\left(\left\|x_{t,j_t}\boldsymbol{W}_{t,j_t:}^\mathsf{T}\cdot\frac{\|\boldsymbol{W}_t\|_F^2}{\|\boldsymbol{W}_{t,j_t:}^\mathsf{T}\|^2}\right\|^2\Big|S\right)\right\}\\ &\le 2B_x^2\cdot\frac{p}{p_0-1}\left\{B_y^2\cdot\frac{q}{q_0}+\mathbb{E}_A(\|\boldsymbol{W}_t\|_F^2\|\boldsymbol{x}_t\|^2|S)\right\}\\ &\le 2B_x^2\cdot\frac{p}{p_0-1}\left(B_y^2\cdot\frac{q}{q_0}+B^2B_x^2\right),\end{aligned} \tag{26}$$

where the second step is due to the triangle inequality, and the third step is due to the construction of $x_{t,j_t}$.

By (25) and (26), taking conditional expectation $\mathbb{E}_A(\cdot|S)$ on both sides of (23) leads to

$$\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^T\mathrm{tr}\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t-\boldsymbol{y}_t)^\mathsf{T}\}\Big|S\right]\le\frac{2B^2}{\eta T}+\frac{\eta B_x^2p}{p_0-1}\left(\frac{B_y^2q}{q_0}+B^2B_x^2\right).$$

□

*Proof of Lemma C.3.* Let $\partial P_\lambda(\boldsymbol{W}_{t+1}^*)$ denote the subgradient of $P_\lambda(\cdot)$ at $\boldsymbol{W}_{t+1}^*$. According to Line 5 of Algorithm 2, we obtain that

$$\boldsymbol{0}\in\boldsymbol{W}_{t+1}^*-\boldsymbol{V}_t+\eta_t\partial P_\lambda(\boldsymbol{W}_{t+1}^*).$$

Hence, by Line 4 of Algorithm 2, there exists a matrix $\boldsymbol{P}_{\lambda,t+1}\in\partial P_\lambda(\boldsymbol{W}_{t+1})$ such that

$$\boldsymbol{W}_{t+1}^*=\boldsymbol{W}_t-\eta_t\breve{\boldsymbol{G}}_t-\eta_t\boldsymbol{P}_{\lambda,t+1}. \tag{27}$$

Due to Line 6 of Algorithm 2 and the condition $\|\boldsymbol{W}^*\|_F\le B$, we can show that $\|\boldsymbol{W}_{t+1}-\boldsymbol{W}^*\|_F\le\|\boldsymbol{W}_{t+1}^*-\boldsymbol{W}^*\|_F$. Combining with (27), we have that

$$\begin{aligned}\|\boldsymbol{W}_{t+1}-\boldsymbol{W}^*\|_F^2 &\le \|\boldsymbol{W}_t-\eta_t\breve{\boldsymbol{G}}_t-\eta_t\boldsymbol{P}_{\lambda,t+1}-\boldsymbol{W}^*\|_F^2\\ &= \|\boldsymbol{W}_t-\boldsymbol{W}^*\|_F^2+\eta_t^2\|\breve{\boldsymbol{G}}_t+\boldsymbol{P}_{\lambda,t+1}\|_F^2\\ &\quad-2\eta_t\mathrm{tr}\left\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}(\breve{\boldsymbol{G}}_t+\boldsymbol{P}_{\lambda,t+1})\right\}.\end{aligned} \tag{28}$$

By the definition of subgradient, it is straightforward to obtain that

$$\mathrm{tr}\left\{(\boldsymbol{W}_{t+1}^* - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} \geq P_\lambda(\boldsymbol{W}_{t+1}^*) - P_\lambda(\boldsymbol{W}^*),$$

and thus,

$$\begin{aligned}
\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} &= \mathrm{tr}\left\{(\boldsymbol{W}_{t+1}^* - \boldsymbol{W}^*)^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} + \mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}_{t+1}^*)^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} \\
&\geq P_\lambda(\boldsymbol{W}_{t+1}^*) - P_\lambda(\boldsymbol{W}^*) + \mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}_{t+1}^*)^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} \\
&= P_\lambda(\boldsymbol{W}_{t+1}^*) - P_\lambda(\boldsymbol{W}^*) + \mathrm{tr}\left\{(\eta_t\breve{\boldsymbol{G}}_t + \eta_t\boldsymbol{P}_{\lambda,t+1})^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} \\
&\geq P_\lambda(\boldsymbol{W}_{t+1}) - P_\lambda(\boldsymbol{W}^*) + \mathrm{tr}\left\{(\eta_t\breve{\boldsymbol{G}}_t + \eta_t\boldsymbol{P}_{\lambda,t+1})^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\}, \qquad (29)
\end{aligned}$$

where the third step is due to (27) and the last step holds since the projection step in Line 6 of Algorithm 2 shrinks the corresponding $P_\lambda(\cdot)$ value.

By (28) and (29), we obtain that

$$\begin{aligned}
&2\eta_t\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\right\} + 2\eta_t\{P_\lambda(\boldsymbol{W}_{t+1}) - P_\lambda(\boldsymbol{W}^*)\} \\
\leq& \|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 - \|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\|_F^2 + \eta_t^2\|\breve{\boldsymbol{G}}_t + \boldsymbol{P}_{\lambda,t+1}\|_F^2 \\
&- 2\eta_t^2\mathrm{tr}\left\{(\breve{\boldsymbol{G}}_t + \boldsymbol{P}_{\lambda,t+1})^\mathsf{T}\boldsymbol{P}_{\lambda,t+1}\right\} \\
=& \|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 - \|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\|_F^2 + \eta_t^2\|\breve{\boldsymbol{G}}_t\|_F^2 - \eta_t^2\|\boldsymbol{P}_{\lambda,t+1}\|_F^2 \\
\leq& \|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 - \|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\|_F^2 + \eta_t^2\|\breve{\boldsymbol{G}}_t\|_F^2. \qquad (30)
\end{aligned}$$

Summing up both side of (30) over $t$ and using the fact that

$$\sum_{t=1}^T P_\lambda(\boldsymbol{W}_{t+1}) \geq \sum_{t=1}^T P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}_1),$$

we obtain that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\left[\mathrm{tr}\left\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\right\} + P_\lambda(\boldsymbol{W}_t) - P_\lambda(\boldsymbol{W}^*)\right] \\
\leq& \frac{1}{T}\sum_{t=1}^T\left(\frac{1}{2\eta_t}\|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 - \frac{1}{2\eta_t}\|\boldsymbol{W}_{t+1} - \boldsymbol{W}^*\|_F^2 + \frac{\eta_t}{2}\|\breve{\boldsymbol{G}}_t\|_F^2\right) + \frac{1}{T}P_\lambda(\boldsymbol{W}_1) \\
=& \frac{1}{2\eta_1 T}\|\boldsymbol{W}_1 - \boldsymbol{W}^*\|_F^2 - \frac{1}{2\eta_{T+1}T}\|\boldsymbol{W}_{T+1} - \boldsymbol{W}^*\|_F^2 \\
&+ \sum_{t=2}^T\left(\frac{1}{2\eta_t T} - \frac{1}{2\eta_{t-1}T}\right)\|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 + \frac{1}{2T}\sum_{t=1}^T\eta_t\|\breve{\boldsymbol{G}}_t\|_F^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1) \\
\leq& 4B^2\left\{\frac{1}{2\eta_1 T} + \sum_{t=2}^T\left(\frac{1}{2\eta_t T} - \frac{1}{2\eta_{t-1}T}\right)\right\} + \frac{1}{2T}\sum_{t=1}^T\eta_t\|\breve{\boldsymbol{G}}_t\|_F^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1) \\
=& \frac{2B^2}{\eta_T T} + \frac{1}{2T}\sum_{t=1}^T\eta_t\|\breve{\boldsymbol{G}}_t\|_F^2 + \frac{1}{T}P_\lambda(\boldsymbol{W}_1), \qquad (31)
\end{aligned}$$

where the third step is because that $\|\boldsymbol{W}_t - \boldsymbol{W}^*\|_F^2 \leq 4B^2$ for all $t$ since $\|\boldsymbol{W}_t\|_F \leq B$ and $\|\boldsymbol{W}^*\|_F \leq B$. Taking the

expectation of both sides over $A$ conditional on $S$ yields that

$$\begin{aligned}
&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\boldsymbol{x}_t(\boldsymbol{W}_t^\mathsf{T}\boldsymbol{x}_t-\boldsymbol{y}_t)\right\}+P_\lambda(\boldsymbol{W}_t)-P_\lambda(\boldsymbol{W}^*)\right]\Bigg|S\right)\\
=&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\right\}+P_\lambda(\boldsymbol{W}_t)-P_\lambda(\boldsymbol{W}^*)\right]\Bigg|S\right)\\
\le&\frac{2B^2}{\eta_T T}+\frac{1}{2T}\sum_{t=1}^{T}\eta_t\mathbb{E}_A(\|\breve{\boldsymbol{G}}_t\|_F^2|S)+\frac{1}{T}P_\lambda(\boldsymbol{W}_1)\\
\le&\frac{2B^2}{\eta_T T}+\left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)B_x^2\cdot\frac{p}{p_0-1}\left(B_y^2\cdot\frac{q}{q_0}+B^2B_x^2\right)+\frac{1}{T}P_\lambda(\boldsymbol{W}_1),
\end{aligned}$$

where we use (25) and (26) to derive the first and third steps, respectively. □

*Proof of Proposition 3.2.* For ease of notation, with a given iteration $t$, we write $\mu_1=\eta_t\lambda_1$ and $\mu_2=\eta_t\lambda_2$. Use $\boldsymbol{x}$ and $\boldsymbol{y}$ to denote $\boldsymbol{W}_{j:}$ and $\boldsymbol{V}_{t,j:}$, respectively, for any fixed $j\in\{1,\dots,p\}$. Since the optimization problem in Line 5 of Algorithm 2 is decomposable, the minimizer of

$$f(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2+\mu_1\|\boldsymbol{x}\|+\mu_2\|\boldsymbol{x}\|_1$$

is the $j$th row of $\boldsymbol{W}_{t+1}^*$.

By the convexity of $f$, suppose $f(\boldsymbol{x})$ is minimized at $\boldsymbol{x}$. Equivalently, we have that

$$\boldsymbol{0}\in\partial f(\boldsymbol{x})=\boldsymbol{x}-\boldsymbol{y}+\mu_1\partial(\|\boldsymbol{x}\|)+\mu_2\partial(\|\boldsymbol{x}\|_1),\tag{32}$$

where $\partial(\|\boldsymbol{x}\|)$ respectively equals $\boldsymbol{x}/\|\boldsymbol{x}\|$ if $\boldsymbol{x}\neq\boldsymbol{0}$, and $\{\boldsymbol{v}\in\mathbb{R}^q:\|\boldsymbol{v}\|\le 1\}$ if $\boldsymbol{x}=\boldsymbol{0}$; and $\partial(\|\boldsymbol{x}\|_1)$ equals $\{(\partial|x_1|,\dots,\partial|x_q|)^\mathsf{T}\}$, with $\partial|x_j|=x_j/|x_j|$ if $x_j\neq 0$ and $\partial|x_j|=[-1,1]$ if $x_j=0$, for $j=1,\dots,q$.

Let $\boldsymbol{y}=(y_1,\dots,y_q)^\mathsf{T}$. Consider the case where $\boldsymbol{x}=0$, i.e., $f$ is minimized at $\boldsymbol{0}$. By (32), there exist $\boldsymbol{v}=(v_1,\dots,v_q)^\mathsf{T}$ satisfying $\|\boldsymbol{v}\|\le 1$ and $\boldsymbol{u}=(u_1,\dots,u_q)^\mathsf{T}$ satisfying $|u_j|\le 1$ for all $j=1.\dots,q$, such that

$$y_j=\mu_1 v_j+\mu_2 u_j\quad\text{ for }\quad j=1,\dots,q.\tag{33}$$

Focus on the minimization of $g(\boldsymbol{u})\triangleq\frac{1}{\mu_1^2}\sum_{j=1}^{q}(y_j-\mu_2 u_j)^2$ over the hypercube $\{\boldsymbol{u}:|u_j|\le 1,j=1,\dots,q\}$, and let $\boldsymbol{u}^*=(u_1^*,\dots,u_q^*)^\mathsf{T}$ denote the minimizer. Clearly, for any $j$, $u_j^*=y_j/\mu_2$ if $|y_j|\le\mu_2$, and $u_j^*=\mathrm{sign}(y_j)$ if $|y_j|>\mu_2$, where $\mathrm{sign}(u)=u/|u|$ for all $u\neq 0$. The existence of decomposition (33) is equivalent to $g(\boldsymbol{u}^*)\le 1$. By definition, $g(\boldsymbol{u}^*)=\|S_{\mu_2}(\boldsymbol{y})\|^2/\mu_1^2$. Hence, (33) is equivalent to $\|S_{\mu_2}(\boldsymbol{y})\|^2\le\mu_1^2$.

Next, we examine the case where $\|S_{\mu_2}(\boldsymbol{y})\|^2>\mu_1^2$, implying $\boldsymbol{x}\neq\boldsymbol{0}$. Define $\mathcal{J}=\{j:x_j=0\}$ and $\mathcal{J}^C=\{j:x_j\neq 0\}$. By (32), for any $j\in\mathcal{J}$, we must have $y_j\in[-\mu_2,\mu_2]$, or equivalently, $S_{\mu_2}(y_j)=0$. For any $j\in\mathcal{J}^C$, (32) gives that

$$y_j=x_j(1+\mu_1/\|\boldsymbol{x}\|+\mu_2/|x_j|),\tag{34}$$

and thus, equivalently,

$$S_{\mu_2}(y_j)=y_j-\mu_2 x_j/|x_j|=x_j(1+\mu_1/\|\boldsymbol{x}\|),$$

where the first step is because $y_j$ and $x_j$ must have the same sign, as indicated by (34). In summary, we have that

$$S_{\mu_2}(y_j)=x_j(1+\mu_1/\|\boldsymbol{x}\|)\text{ for all }j=1,\dots,q.\tag{35}$$

Taking the Euclidian norm over both sides of (35) yields that

$$\|S_{\mu_2}(\boldsymbol{y})\|=\|\boldsymbol{x}\|+\mu_1.$$

Therefore, replacing $\|\boldsymbol{x}\|$ in (35) with $\|S_{\mu_2}(\boldsymbol{y})\|-\mu_1$, we obtain that

$$\boldsymbol{x}=(\|S_{\mu_2}(\boldsymbol{y})\|-\mu_1)\cdot\frac{S_{\mu_2}(\boldsymbol{y})}{\|S_{\mu_2}(\boldsymbol{y})\|}.\tag{36}$$

To sum up, if $\boldsymbol{x}$ is the minimizer of $f$, then $\boldsymbol{x} = \mathbf{0}$ if and only if $\|S_{\mu_2}(\boldsymbol{y})\|^2 \le \mu_1^2$. Otherwise, $\boldsymbol{x}$ has the form (36). Consequently, we obtain that

$$\boldsymbol{x} = (\|S_{\mu_2}(\boldsymbol{y})\| - \mu_1)_+ \cdot \frac{S_{\mu_2}(\boldsymbol{y})}{\|S_{\mu_2}(\boldsymbol{y})\|},$$

where we use the convention $0/0 = 0$. □

*Proof of Lemma C.4.* The proof for Lemma C.4 is analogous to the proof of Lemma C.2. Specifically, the inequality (23) still holds. It suffices to provide an upper bound for $\mathbb{E}_A(\|\breve{\boldsymbol{G}}_t\|_F^2|S)$. By the construction of $\tilde{\boldsymbol{x}}_t$ and $\breve{\boldsymbol{G}}_t$ in Algorithm 3, we have that

$$\mathbb{E}_A(\tilde{\boldsymbol{x}}_t|S) = \frac{p_0}{p} \cdot \boldsymbol{x}_t, \quad \mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t\|^2|S) = \frac{p_0}{p} \cdot \|\boldsymbol{x}_t\|^2, \tag{37}$$

and $\|\breve{\boldsymbol{G}}_t\|_F^2 = \frac{q^2}{q_0^2}\|\tilde{\boldsymbol{x}}_t\|^2\|\boldsymbol{\phi}_t\|^2 \le \frac{q^2}{q_0}\|\tilde{\boldsymbol{x}}_t\|^2$. Therefore, we obtain that

$$\mathbb{E}_A(\|\breve{\boldsymbol{G}}_t\|_F^2|S) \le \frac{q^2 p_0}{p q_0} \cdot \|\boldsymbol{x}_t\|^2 \le \frac{q^2 p_0}{p q_0} B_x^2, \tag{38}$$

which by (23), leads to

$$\mathbb{E}_A\left[\frac{1}{T}\sum_{t=1}^{T} \mathrm{tr}\{(\boldsymbol{W}_t - \boldsymbol{W}^*)^\mathsf{T}\breve{\boldsymbol{G}}_t\}\Big|S\right] \le \frac{2B^2}{\eta T} + \frac{\eta}{2} \cdot \frac{q^2 p_0}{p q_0} B_x^2.$$

□

*Proof of Lemma C.5.* By the construction of $\mathcal{O}_t$ in Algorithm 3, for any constant vector $\boldsymbol{b} \in \mathbb{R}^q$, we have that

$$\mathbb{E}(\|\boldsymbol{b}_{\mathcal{O}_t}\|_1) = \frac{q_0}{q}\|\boldsymbol{b}\|_1.$$

By the law of iterative expectations, we then have that, for any fixed $\boldsymbol{W}$,

$$\tilde{L}_t(\boldsymbol{W}) = \frac{q}{q_0}\mathbb{E}_{A_t}\{\mathbb{E}_{A_t}(\|\boldsymbol{W}_{:\mathcal{O}_t}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_{t,\mathcal{O}_t}\|_1|\tilde{\boldsymbol{x}}_t, S_t)|S_t\} = \mathbb{E}_{A_t}(\|\boldsymbol{W}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_t\|_1|S_t).$$

In addition, by Jensen's inequality and (37),

$$\begin{aligned}
\mathbb{E}_{A_t}(\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\||S_t) &= \mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\||S) \\
&\le \left\{\mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\|^2|S)\right\}^{1/2} \\
&= \left\{\mathbb{E}_A(\|\tilde{\boldsymbol{x}}_t\|^2|S) + \|\boldsymbol{x}_t\|^2 - 2\boldsymbol{x}_t^\mathsf{T}\mathbb{E}_A(\tilde{\boldsymbol{x}}_t|S)\right\}^{1/2} \\
&= \|\boldsymbol{x}_t\|\sqrt{1 - \frac{p_0}{p}},
\end{aligned}$$

where the first step is because that the randomness of $\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t$ comes from $S_t$ and $A_t$ only. Therefore, for any matrix $\boldsymbol{W}$ and any $t$, we arrive at

$$\begin{aligned}
|\tilde{L}_t(\boldsymbol{W}) - L_t(\boldsymbol{W})| &= \left|\mathbb{E}_{A_t}(\|\boldsymbol{W}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_t\|_1 - \|\boldsymbol{W}^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t\|_1\big|S_t)\right| \\
&\le \mathbb{E}_{A_t}\left\{\left|\|\boldsymbol{W}^\mathsf{T}\tilde{\boldsymbol{x}}_t - \boldsymbol{y}_t\|_1 - \|\boldsymbol{W}^\mathsf{T}\boldsymbol{x}_t - \boldsymbol{y}_t\|_1\right|\big|S_t\right\} \\
&\le \mathbb{E}_{A_t}\left\{\|\boldsymbol{W}^\mathsf{T}(\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t)\|_1\big|S_t\right\} \\
&\le \sqrt{q}\,\mathbb{E}_{A_t}\left\{\|\boldsymbol{W}^\mathsf{T}(\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t)\|\big|S_t\right\} \\
&\le \sqrt{q}\,B\|\boldsymbol{x}_t\|\sqrt{1 - \frac{p_0}{p}},
\end{aligned}$$

where the second step is due to Jensen's inequality, the third step comes from the triangle inequality of the $L_1$ norm, the fourth step follows from the fact that $\|\boldsymbol{u}\|_1 \le \sqrt{q}\|\boldsymbol{u}\|$ for any $\boldsymbol{u} \in \mathbb{R}^q$, and the last step is due to $\|\boldsymbol{A}\boldsymbol{u}\| \le \|\boldsymbol{A}\|_F\|\boldsymbol{u}\|$ for any matrix $\boldsymbol{A}$ and vector $\boldsymbol{u}$ with suitable dimensions. The proof is then completed due to the upper bound of $\|\boldsymbol{x}_t\|$. □

*Proof of Lemma C.6.* Based on the proof of Lemma C.3, the proof of Lemma C.6 is straightforward. Indeed, results (27)-(30) still hold for Algorithm 4. Hence, identical to (31), we have that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\intercal\breve{\boldsymbol{G}}_t\right\}+P_\lambda(\boldsymbol{W}_t)-P_\lambda(\boldsymbol{W}^*)\right]\\
\leq&\frac{2B^2}{\eta_T T}+\frac{1}{2T}\sum_{t=1}^{T}\eta_t\|\breve{\boldsymbol{G}}_t\|_F^2+\frac{1}{T}P_\lambda(\boldsymbol{W}_1),
\end{aligned}$$

which yields that

$$\begin{aligned}
&\mathbb{E}_A\left(\frac{1}{T}\sum_{t=1}^{T}\left[\mathrm{tr}\left\{(\boldsymbol{W}_t-\boldsymbol{W}^*)^\intercal\breve{\boldsymbol{G}}_t\right\}+P_\lambda(\boldsymbol{W}_t)-P_\lambda(\boldsymbol{W}^*)\right]\Bigg|S\right)\\
\leq&\frac{2B^2}{\eta_T T}+\frac{1}{2T}\sum_{t=1}^{T}\eta_t\mathbb{E}_A(\|\breve{\boldsymbol{G}}_t\|_F^2|S)+\frac{1}{T}P_\lambda(\boldsymbol{W}_1)\\
\leq&\frac{2B^2}{\eta_T T}+\left(\frac{1}{T}\sum_{t=1}^{T}\eta_t\right)\frac{q^2p_0}{2pq_0}B_x^2+\frac{1}{T}P_\lambda(\boldsymbol{W}_1),
\end{aligned}$$

where the second step follows from (38). □