

# High-Dimensional Tail Index Regression

## with An Application to Text Analyses of Viral Posts in Social Media\*

Yuya Sasaki<sup>†</sup> Jing Tao<sup>‡</sup> Yulong Wang<sup>§</sup>

### Abstract

Motivated by the empirical observation of power-law distributions in the credits (e.g., “likes”) of viral social media posts, we introduce a high-dimensional tail index regression model and propose methods for estimation and inference of its parameters. First, we present a regularized estimator, establish its consistency, and derive its convergence rate. Second, we introduce a debiasing technique for the regularized estimator to facilitate inference and prove its asymptotic normality. Third, we extend our approach to handle large-scale online streaming data using stochastic gradient descent. Simulation studies corroborate our theoretical findings. We apply these methods to the text analysis of viral posts on X (formerly Twitter) related to LGBTQ+ topics.

**Keywords:** high-dimensional data, social media, tail index, Pareto, text data analysis.

---

\*First arXiv date: Mar 02 2024. We thank Yuan Liao, Denis Chetverikov, and participates at the ESIF Economics and AI+ML Meeting for very helpful comments and suggestions.

<sup>†</sup> Yuya Sasaki: [yuya.sasaki@vanderbilt.edu](mailto:yuya.sasaki@vanderbilt.edu). Brian and Charlotte Grove Chair and Professor of Economics, Department of Economics, Vanderbilt University, VU Station B #351819, 2301 Vanderbilt Place, Nashville, TN 37235-1819.

<sup>‡</sup> Jing Tao: [jingtao@uw.edu](mailto:jingtao@uw.edu). Assistant Professor of Economics, Department of Economics, University of Washington, Box 353330, Savery 305, University of Washington, Seattle, WA 98195-3330.

<sup>§</sup> Yulong Wang: [ywang402@maxwell.syr.edu](mailto:ywang402@maxwell.syr.edu). Associate Professor of Economics, Department of Economics, Syracuse University, 110 Eggers Hall, Syracuse, NY 13244-1020.

# 1 Introduction

A large literature is dedicated to tail features of distributions – see de Haan and Ferreira (2007) and Resnick (2007) for reviews and references. A distribution  $F$  regularly varies with exponent  $\alpha$  if its tail is well approximated by a Pareto distribution with shape parameter  $\alpha$ . This regularity condition implies that common tail features of interest, such as tail probabilities, extreme quantiles, and tail conditional expectations, can be expressed in terms of  $\alpha$ . Estimates of these features are obtained by plugging in estimated values of  $\alpha$ . The literature contains numerous suggestions along these lines, some of which are reviewed in the surveys cited below.

Consider the distribution of credits in social media to motivate this framework in contemporary applications. Figure 1 displays the so-called log-log plot for the distribution of the number  $Y$  of “likes” in LGBTQ+ posts in X (formerly Twitter). If the distribution of  $Y$  is Pareto with exponent  $\alpha$ , the log-log plot would appear linear, as in this figure, with its slope indicating  $-1/\alpha$ . This observation motivates us to use the aforementioned technology for the analyses of viral posts on social media.

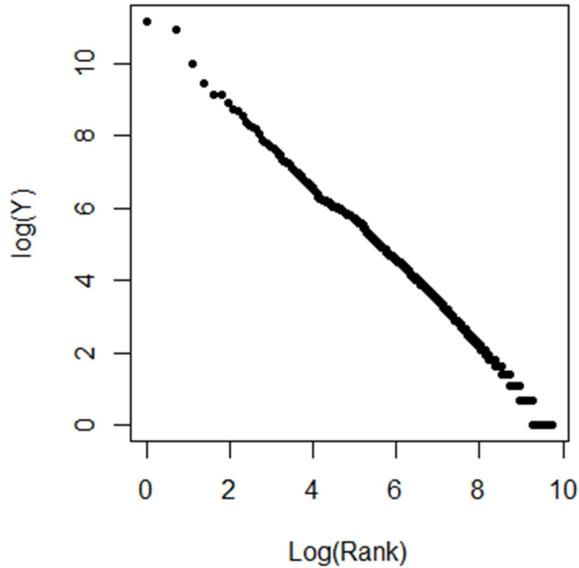


Figure 1: The log-log plot for the distribution of “likes” in posts about LGBTQ+ in X. The horizontal axis plots the rank of  $Y$  while the vertical axis plots  $\log(Y)$ .

We emphasize that the Pareto tail is not unique to our dataset. It has also been documented and explained for numerous economics, finance, and insurance datasets, including city sizes, firm sizes, stock returns, and natural disasters. For examples, see Gabaix (2009, 2016). In text analysis and linguistics, the so-called Zipf’s law states that given a large sample of words used, the frequency of any word is inversely proportional to its rank in the frequency table. This empirical finding can also be characterized by the Pareto distribution with  $\alpha \approx 1$  (e.g., Fagan and Gençay, 2010, p.139).

Suppose that the conditional distribution of  $Y$  given  $X$  has an approximately Pareto tail with shape parameter  $\alpha(X)$  depending on covariates  $X$ . We are interested in the effect of  $X$  on the tail features of  $Y$  through  $\alpha(X)$ . One family of existing methods imposes a parametric structure on  $\alpha(X)$ . Wang and Tsai (2009) propose a tail index regression (TIR) method by modeling  $\alpha(X) = \exp(X^\top \theta_0)$  and estimating the pseudo-parameter  $\theta_0$ . Nicolau, Rodrigues, and Stoykov (2023) extend the TIR method to accommodate weakly dependent data. Li, Leng, and You (2020) consider the semiparametric setup  $\alpha(X) = \alpha(X_1, X_2) = \exp(X_1^\top \theta_0 + \eta(X_2))$  for some smooth function  $\eta$ . By combining  $\alpha(X) = \exp(X^\top \theta_0)$  with a power transformation of  $Y$ , Wang and Li (2013) study the estimation of conditional extreme quantiles of  $Y$  given  $X$ . Another family of existing methods considers fully nonparametric models and local smoothing (e.g., Gardes and Girard, 2010; Gardes, Guillou, and Schorgen, 2012; Daouia, Gardes, Girard, and Lekina, 2010; Daouia, Gardes, and Girard, 2013)).

Common in all of these existing approaches is that  $X$  is assumed to be of a fixed and low dimension. In this paper, we aim to relax this restriction by allowing the dimension of  $X$  to increase with the sample size and potentially exceed the sample size. Our empirical question related to Figure 1 motivates this high-dimensional model. Specifically, let  $Y_i$  denote the number of “likes” of the  $i$ -th post, and let  $X_i$  denote a long vector of binary indicators of whether this post contains a list of keywords. Smaller values of  $\alpha(X)$  imply that using the words indicated by such  $X$  entails more extreme numbers of “likes.” Essentially, we are asking how to write viral posts. A high-dimensional setup is crucial since the number of keywords is potentially huge.

To address this question, we develop a novel high-dimensional tail index regression (HDTIR) method. Specifically, by modifying the TIR method (Wang and Tsai, 2009), we propose an  $L^1$ -regularized maximum likelihood estimator. For inference, we further propose debiasing the regularized estimator and establishing its asymptotic normality. Two alternative methods are provided for

debiased estimation and inference: one based on sample splitting and the other based on cross-fitting.

Furthermore, to accommodate a large stream of text data from successive posts in social media, we extend our method to handle large-scale online and streaming datasets by employing a variation of the stochastic gradient descent algorithm, an active research area. Building on recent developments from Agarwal, Negahban, and Wainwright (2012), Chen, Lee, Tong, and Zhang (2020), Han, Luo, Lin, and Huang (2024), and others, we provide an algorithm for implementing our HDTIR method and derive its asymptotic normality. Our approach has two distinctive features. First, our stochastic gradient (score function) is sub-exponential rather than sub-Gaussian. Second, while existing literature predominantly focuses on algorithms for high-dimensional linear regression models, our work addresses a high-dimensional nonlinear tail index model.

Additionally, we highlight two technical differences between our approach and the extensive literature on high-dimensional methods available today. First, our analysis focuses on the tail of the underlying distribution by using only the tail observations  $Y > w_n$  for a large threshold  $w_n$ . This threshold acts as a tuning parameter, similar to the bandwidth in nonparametric regression. We impose restrictions on  $w_n$  to mitigate its impact in the asymptotic analysis, similar to undersmoothing bandwidth. Second, the convergence rate in our setting is determined by the *tail* sample size  $n_0 = \sum_{i=1}^n 1[Y_i > w_n]$  instead of the full sample size  $n$ . Since  $n_0$  is much smaller than  $n$ , the convergence rate of our HDTIR is slower than that of existing methods that use the full sample.

To the best of our knowledge, the method proposed in the current paper is the first systematic study of estimation and inference theory for the high dimensional tail index regression model. This constitutes the key contribution of the current paper. The estimation and inference problem in the high-dimensional tail regression model is related to the extensive literature on high-dimensional generalized linear models (e.g., van de Geer (2008), Negahban, Yu, Wainwright, and Ravikumar (2009), Huang and Zhang (2012), van de Geer, Bühlmann, Ritov, and Dezeure (2014), Zhang and Zhang (2014), Belloni, Chernozhukov, Chetverikov, and Wei (2018), Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018), Cai, Guo, and Ma (2023), among many others). However, none of the aforementioned papers focuses on tail index regression. Our work also aims to contribute to the vast literature of text analysis.  $L^1$ -regularized estimation has been extensively applied to high-dimensional text regressions (e.g., Taddy, 2013). Nonetheless, there is no method tailored to analyzing

tail features of distributions of credits, such as the number of “likes” for viral posts in social media. Our proposal addresses this gap in the literature as well.

The current paper is also related to the recent literature on shrinkage methods with heavy tail data (e.g., Wong, Li, and Tewari, 2020; Fan, Wang, and Zhu, 2021; Zhu and Zhou, 2021; Babii, Ball, Ghysels, and Striaukas, 2023). These methods focus on modeling the conditional mean  $\mathbb{E}[Y|X]$  using all  $n$  observations. The heavy tail feature typically leads to a slower convergence rate, i.e. from  $\log p$  to a polynomial of  $p$ , where  $p$  denotes the dimension of  $X$ . The asymptotic distributions become more complicated, as does the subsequent statistical inference. In contrast, our method relies on the regular variation assumption and focuses on the conditional tail index of  $Y$ . This tail feature requires using only the tail  $n_0 < n$  observations, but restores the conventional  $\log p$  rate. Note that our  $p$  may increase with  $n_0$ , and we leave its relation with  $n$  unspecified.

The rest of the paper is organized as follows. Section 2 presents the method and its theory of HDTIR, and Section 3 extends the analysis to conditional quantiles. Section 4 further extends our method to large-scale online data. Monte Carlo simulations in Section 5 demonstrate that the proposed HDTIR has excellent small-sample performance. We apply the method to text analyses of viral posts in X in Section 6. Mathematical proofs and technical details are relegated to the appendix.

Throughout the paper, we use the following notation. For a  $p$ -dimensional vector  $X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ , we use  $\|X\|_q = (\sum_{i=1}^p |X_i|^q)^{1/q}$  to denote the vector  $\ell_q$  norm for  $1 \leq q < \infty$ , and  $\|X\|_\infty = \max_{1 \leq i \leq q} |X_i|$  to denote the vector maximum norm. For a set  $S \subseteq \{1, \dots, p\}$ , let  $X_S = \{X_j : j \in S\}$  and  $S^c$  be the complement of  $S$ . For a  $p \times q$  matrix  $A = (a_{i_1 i_2}) \in \mathbb{R}^{p \times q}$ , we use  $\|A\|_1 = \sum_{i_1=1}^p \sum_{i_2=1}^q |a_{i_1 i_2}|$ ,  $\|A\|_2 = \|A\|_F = \{\sum_{i_1=1}^p \sum_{i_2=1}^q (a_{i_1 i_2})^2\}^{1/2}$  and  $\|A\|_\infty = \max_{1 \leq i_1 \leq p, 1 \leq i_2 \leq q} |a_{i_1 i_2}|$  to denote the element-wise  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$ , respectively. Let  $\|A\|_{\ell_d} = \sup_{X \in \mathbb{R}^q, \|X\|_d \leq 1} |AX|_d$  denote the matrix operator norm for  $1 \leq d \leq \infty$ . More specifically, the operator  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  norms are denoted by  $\|A\|_{\ell_1} = \max_{1 \leq i_2 \leq q} \sum_{i_1=1}^p |a_{i_1 i_2}|$ ,  $\|A\|_{\ell_2} = \max_{1 \leq i_2 \leq q} \{\sum_{i_1=1}^p (a_{i_1 i_2})^2\}^{1/2}$  and  $\|A\|_{\ell_\infty} = \max_{1 \leq i_1 \leq p} \sum_{i_2=1}^q |a_{i_1 i_2}|$ , respectively. Let  $I_p$  be the  $p \times p$  identity matrix. For two positive sequences  $\{a_n\}$  and  $\{b_n\}$ ,  $a_n \gtrsim b_n$  means that  $a_n > cb_n$  for all  $n$  large enough and some constant  $c$ ,  $a_n \lesssim b_n$  if  $b_n \gtrsim a_n$  holds, and  $a_n \asymp b_n$  means that  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ . Moreover,  $a_n \ll b_n$  means  $a_n/b_n \rightarrow 0$ . For any random variables  $X_1, \dots, X_n$  and functions  $h(\cdot)$ , let  $\mathbb{E}_n\{h(U_i)\} = \sum_{i=1}^n h(U_i)/n$  be the empirical average of  $\{h(U_i)\}_{i=1}^n$ . Let  $\dot{h}(\cdot)$  and  $\ddot{h}(\cdot)$  be the first and second-order derivatives of

a univariate function, and let  $\nabla$  denote the operator for gradient or subgradient.

## 2 High Dimensional Tail Index Regression

### 2.1 Regularized Estimation

Let  $\{(X_i, Y_i)\}_{i=1}^n$  be  $n$  copies of  $\{Y, X\}$ , where  $Y$  is a real-valued response of interest and  $X$  is a  $p$ -dimensional random vector of explanatory factors with possibly  $p = p_n \rightarrow \infty$  as the sample size  $n$  diverges to  $\infty$ . We are interested in modeling the effect of  $X$  on the tail feature of the distribution of  $Y$ . Without loss of generality, we focus on the right tail and collect observations  $Y_i > w_n$  for some  $w_n$ . Let  $n_0 := \sum_{i=1}^n \mathbb{1}[Y_i \geq w_n]$  denote the effective sample size, and rearrange the indices such that  $\mathbb{1}[Y_i \geq w_n] = 1$  for all  $i \in \{1, \dots, n_0\}$ . The following assumptions describe our model.

**Assumption 1** (Conditional Pareto Tail). *For  $t > 0$ ,*

$$\mathbb{P}(Y > t | X = x) = t^{-\alpha(X)} \mathcal{L}(t; x)$$

with  $\alpha(X) = \exp(X^\top \theta_0)$ , where  $\mathcal{L}(t; x)$  satisfies

$$\mathcal{L}(t; x) = c_0(x) + c_1(x) t^{-\beta(x)} + r(t, x) \text{ as } t \rightarrow \infty$$

for some functions  $\beta(x) > \underline{\beta} > 0$ ,  $c_0(x) > 0$ ,  $c_1(x) \in \mathbb{R}$ , and  $r(t, x)$  such that

$$\sup_x r(t, x) t^{\beta(x)} \rightarrow 0 \text{ as } t \rightarrow \infty.$$

**Assumption 2** (Compact Support).  $X_i$  is i.i.d. For each  $j = 1, \dots, p$ ,  $X_{i,j}$  has a compact support  $\mathcal{X}_j$  with  $\sup_{x \in \mathcal{X}_j} f_{X_{i,j}}(x) < \bar{f} < \infty$ .

Assumption 1 imposes the restriction that the tail of the conditional distribution of  $Y$  given  $X$  can be approximated by Pareto with exponent  $\alpha(X)$  (e.g., Wang and Tsai, 2009). The first-order term  $\mathcal{L}(t; x)$  is slowly varying in the sense that  $\mathcal{L}(ty; x)/\mathcal{L}(t; x) \rightarrow 1$  as  $t \rightarrow \infty$  for any  $x$  and  $y > 0$ . The higher-order term  $\beta(X)$  characterizes the deviation from the exact Pareto, which substantially complicates the analysis. This condition with a constant  $X$  has been widely imposed and studied in the existing literature (e.g., Hall, 1982; de Haan and Ferreira, 2007). We remark that the Pareto tail approximation holds for many commonly used heavy-tailed distributions such as Student-t, F,

Cauchy distributions. Specifically, if  $Y$  given  $X = x$  is Student-t distributed with degree of freedom  $\nu(x)$ , Assumption 1 is then satisfied with  $\alpha(x) = \nu(x)$  and  $\beta(x) = 2$ .

Assumption 2 imposes that each coordinate of  $X_i$  has a compact support. This is coherent with our empirical application, in which  $X_i$  is a vector of binary indicators of keywords. We can also relax the i.i.d. condition at the cost of more sophisticated theory, but we focus on this sampling assumption to explicate our main contribution concerning the high-dimensional setup.

We now introduce our high-dimensional tail index regression (HDTIR) estimator. Define the negative log-likelihood function of  $Y$  conditional on  $Y \geq w_n$  by

$$\ell_{n_0}(\theta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \{\exp(X_i^\top \theta) \log(Y_i/w_n) - X_i^\top \theta\}. \quad (2.1)$$

Our regularized HDTIR estimator is given by

$$\hat{\theta} = \arg \min_{\theta} \{\ell_{n_0}(\theta) + \lambda_{n_0} \|\theta\|_1\}. \quad (2.2)$$

We denote the sparsity level of the parameter as  $s_0$ , i.e.,  $\|\theta\|_0 \leq s_0$ . To derive asymptotic properties of  $\hat{\theta}$ , we introduce the following undersmoothing condition on  $w_n$ .

**Assumption 3** (Undersmoothing).  $w_n \rightarrow \infty$  satisfies

$$(i) \quad \frac{\log p}{n \mathbb{E} [c_0(X_i) w_n^{-\alpha(X_i)}]} \rightarrow 0 \quad \text{and}$$

$$(ii) \quad s_0 w_n^{-2\beta} \frac{\log p}{n \mathbb{E} [c_0(X_i) w_n^{-\alpha(X_i)}]} \rightarrow 0,$$

where  $\underline{\beta} > 0$  and  $c_0(x) > 0$  are defined in Assumption 1.

Assumption 3 imposes restrictions about the threshold  $w_n$ . Part (i) prohibits  $w_n$  from growing too fast with  $n$ . This condition comes from the requirement that  $n_0/\log p \rightarrow \infty$  and the fact (Wang and Tsai, 2009, Theorem 1) that

$$\frac{n_0}{n} = \mathbb{E} [c_0(X_i) w_n^{-\alpha(X_i)}] (1 + o_p(1)).$$

It guarantees sufficient tail observations relative to  $p$ . In contrast, Part (ii) prohibits  $w_n$  from growing too slowly relative to the sparsity  $s_0$ . It is close in spirit to the undersmoothing choice of bandwidth in

nonparametric estimation. Otherwise, the Pareto approximation incurs a non-negligible asymptotic bias, which involves the second-order parameter  $\underline{\beta}$  (and even  $\beta(x)$ ). These second-order parameters are difficult to estimate and hence bias correction relying on these second-order parameters is infeasible.

Let  $\Sigma_{w_n} = \mathbb{E}[X_i X_i^\top | Y_i > w_n]$  and  $Z_{ni} = \Sigma_{w_n}^{-1/2} X_i$ . The following theorem establishes the consistency and the convergence rate of this regularized estimator.

**Theorem 1.** *Suppose that Assumptions 1-3 hold. Suppose also that, for constants  $C_1 > 0$  and  $C_2 > 1$  independent of  $n$ ,  $p$ , and  $w_n$ ,  $\|\theta_0\|_2 \leq C_1$ , and  $C_2^{-1} \leq \lambda_{\min}(\Sigma_{w_n}) \leq \lambda_{\max}(\Sigma_{w_n}) \leq C_2$  hold. Let  $\lambda_{n_0} = c\sqrt{(\log p)/n_0}$  for some constant  $c > 0$ . If  $s_0 \lesssim n_0/(\log p)$ , then*

$$\|\hat{\theta} - \theta_0\|_1 \lesssim \sqrt{\frac{s_0^2(\log p)}{n_0}}, \quad \|\hat{\theta} - \theta_0\|_2 \lesssim \sqrt{\frac{s_0(\log p)}{n_0}}, \quad \text{and} \quad \frac{1}{n_0} \sum_{i=1}^{n_0} [X_i^\top (\hat{\theta} - \theta_0)]^2 \lesssim \frac{s_0(\log p)}{n_0}$$

hold with probability approaching one.

Theorem 1 establishes the rate of convergence for the proposed regularized estimator under mild conditions. This result extends previous work on generalized linear models with canonical links (e.g., Negahban et al., 2009; Gardes and Girard, 2010) and those focusing on generalized linear models with binary outcomes (e.g., van de Geer, 2008; Cai et al., 2023). However, as extensively discussed in the literature,  $\hat{\theta}$  cannot be directly used to construct a confidence interval for  $\theta_0$ . In the next section, we introduce a debiased estimator to address this issue and facilitate statistical inference.

## 2.2 Debiased Estimation and Inference

The regularized estimator generally entails non-negligible regularization biases relative to its sampling variations and cannot be directly used for statistical inference based on its asymptotic distribution. In this light, the current section proposes two approaches to debiased estimation and inference. One approach is based on sample splitting, and the other is based on cross-fitting.

Note that the score and Hessian of  $\ell_{n_0}$  evaluated at  $\hat{\theta}$  take the forms of

$$\begin{aligned} \dot{\ell}_{n_0}(\hat{\theta}) &= \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \exp(X_i^\top \hat{\theta}) \log(Y_i/w_n) - 1 \right\} X_i \quad \text{and} \\ \ddot{\ell}_{n_0}(\hat{\theta}) &= \frac{1}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_n) \exp(X_i^\top \hat{\theta}) X_i X_i^\top, \end{aligned}$$

respectively.

### 2.2.1 Sample Splitting

We split the samples so that the initial estimation and bias correction steps are conducted on independent datasets. Without loss of generality, the effective sample of size  $2n_0$  is randomly divided into two disjoint subsets  $\mathcal{D}_1 = \{(X_i, Y_i)\}_{i=1}^{n_0}$  and  $\mathcal{D}_2 = \{(X_i, Y_i)\}_{i=n_0+1}^{2n_0}$ . We use  $\mathcal{D}_2$  to obtain  $\widehat{\theta}$  via (2.2) and use  $\{X_i\}_{i=1}^{n_0}$  in the subsample  $\mathcal{D}_1$  for the debiasing step described below. Let

$$\widehat{u}_j = \arg \min_{u \in \mathbb{R}^p} u^\top \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} X_i X_i^\top \right] u \quad (2.3)$$

$$s.t. \quad \left\| \frac{1}{n_0} \sum_{i=1}^{n_0} X_i X_i^\top u - e_j \right\|_\infty \leq \gamma_{1n_0} \quad \text{and} \quad (2.4)$$

$$\max_{1 \leq i \leq n_0} |X_i^\top u| \leq \gamma_{2n_0}, \quad (2.5)$$

where  $\{e_j\}_{j=1}^p$  denotes the canonical basis of the Euclidean space  $\mathbb{R}^p$ , and  $\gamma_{1n_0}$  and  $\gamma_{2n_0}$  satisfy the conditions stated in Assumption 4 below.

**Assumption 4.** For some constants  $c, c', c'' > 0$ , (i)  $\lambda_{n_0} = c\sqrt{(\log p)/n_0}$ , (ii)  $\gamma_{1n_0} = c'\sqrt{(\log p)/n_0}$ , and (iii)  $\gamma_{2n_0} = c''\sqrt{\log n_0}$ .

For each coordinate  $j = 1, \dots, p$ , the debiased estimator is defined by

$$\widetilde{\theta}_j := \widehat{\theta}_j - \frac{\widehat{u}_j^\top}{n_0} \sum_{i=1}^{n_0} \left\{ \exp(X_i^\top \widehat{\theta}) \log(Y_i/w_n) - 1 \right\} X_i,$$

where  $\widehat{u}_j \in \mathbb{R}^p$  is the projection direction constructed by (2.3)–(2.5) using the subsample  $\mathcal{D}_1$ , while  $\widehat{\theta}$  derives from (2.2) using the subsample  $\mathcal{D}_2$ .

We emphasize that the construction of  $\widehat{u}_j$  takes advantage of our Pareto tail approximation. Specifically, the existing literature (e.g., Cai et al., 2023, eq.(11)) constructs  $\widehat{u}_j$  using the Hessian, which does not involve  $Y_i$ . By conditioning on  $X_i$  and constructing  $\widehat{\theta}$  using a different subsample, the debias term maintains conditional zero mean. However, our Hessian involves  $\{\exp(X_i^\top \widehat{\theta}) \log(Y_i/w_n)\} X_i X_i^\top$ , which contains  $Y_i$ . To address this issue, we note that the conditional distribution of  $\log(Y_i/w_n)$  given  $X_i$  is approximately exponential with parameter  $\exp(X_i^\top \theta_0)$ . It follows that

$$\mathbb{E} \left[ \exp(X_i^\top \widehat{\theta}) \log(Y_i/w_n) \right] X_i X_i^\top \approx \mathbb{E} [X_i X_i^\top],$$

which motivates our construction of  $\widehat{u}_j$  in (2.3).

Define the variance estimator by

$$\widehat{V}_{1j} := \widehat{u}_j^\top \left[ \frac{1}{n_0} \sum_{i=1}^{n_0} X_i X_i^\top \right] \widehat{u}_j. \quad (2.6)$$

We now formally establish the asymptotic property for this debiased estimator,  $\widetilde{\theta}_j$ .

**Theorem 2.** *Suppose that Assumptions 1-4 are satisfied. If  $s_0 \ll \frac{\sqrt{n_0}}{\log p \sqrt{\log n_0}}$ , then*

$$\sqrt{n_0} \widehat{V}_{1j}^{-1/2} (\widetilde{\theta}_j - \theta_{0j}) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n_0 \rightarrow \infty.$$

One can obtain similar result in Theorem 1 without sampling splitting. However, following the insight from Cai et al. (2023), with sample splitting, the debiased estimator achieves asymptotic normality without requiring the inverse matrix  $\left[ \frac{1}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_n) \exp(X_i^\top \widehat{\theta}) X_i X_i^\top \right]$  to be weakly sparse, which relaxes a standard assumption in the literature (see, e.g., van de Geer et al. (2014), Javanmard and Montanari (2014) and Zhang and Zhang (2014) for linear regression models). We propose a more general cross-fitting method in next section.

### 2.2.2 Cross Fitting

The sample splitting approach introduced in Section 2.2.1 uses only half of the sample. To overcome this deficiency, we now propose a cross-fitting approach.

Take a  $K$ -fold random partition  $(\mathcal{D}_k)_{k=1}^K$  of the indices  $[n_0] = \{1, \dots, n_0\}$  so that the size of each fold  $\mathcal{D}_k$  is  $n_k = n_0/K$ . For each  $k = 1, \dots, K$ , define the set  $\mathcal{D}_k^c = \{1, \dots, n_0\} \setminus \mathcal{D}_k$  of indices for the complement of the fold. In practice,  $K$  can be a constant as small as 2.

For each  $k \in \{1, \dots, K\}$ , we estimate  $\widehat{\theta}_k$  via (2.2) by using the subsample of  $\mathcal{D}_k^c$ , and estimate  $\widehat{u}_{j,k}$  via (2.3)–(2.5) by using the subsample  $\mathcal{D}_k$ . Specifically, consider the following algorithm to obtain  $\widehat{u}_{j,k}$  with cross-fitting:

$$\widehat{u}_{j,k} = \arg \min u^\top \left( \frac{1}{n_0 - n_k} \sum_{i \in \mathcal{D}_k} X_i X_i^\top \right) u \quad (2.7)$$

$$\text{s.t. } \left\| \left( \frac{1}{n_0 - n_k} \sum_{i \in \mathcal{D}_k} X_i X_i^\top \right) u - e_j \right\|_\infty \leq \gamma_{1n_0} \text{ and} \quad (2.8)$$

$$\max_{i \in \mathcal{D}_k} |X_i^\top u| \leq \gamma_{2n_0}. \quad (2.9)$$

Then, for each  $j \in \{1, \dots, p\}$ , let

$$\tilde{\theta}_{j,k} := \hat{\theta}_{j,k} - \frac{\hat{u}_{j,k}^\top}{n_k} \sum_{i \in \mathcal{D}_k} \left\{ \exp(X_i^\top \hat{\theta}_k) \log(Y_i/w_n) - 1 \right\} X_i,$$

and define the debiased estimator by taking the average

$$\tilde{\theta}_j := \frac{1}{K} \sum_{k=1}^K \tilde{\theta}_{j,k}. \quad (2.10)$$

Letting its asymptotic variance estimator be defined by

$$\hat{V}_{2j} := \frac{1}{K} \sum_{k=1}^K \left[ \hat{u}_{j,k}^\top \left( \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} X_i X_i^\top \right) \hat{u}_{j,k} \right],$$

we now establish the asymptotic property for the debiased estimator  $\tilde{\theta}_j$ .

**Theorem 3.** *Suppose that Assumptions 1-4 are satisfied. If  $s_0 \ll \frac{\sqrt{n_0}}{\log p \sqrt{\log n_0}}$ , then*

$$\sqrt{n_0} \hat{V}_{2j}^{-1/2} (\tilde{\theta}_j - \theta_{0j}) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n_0 \rightarrow \infty.$$

Theorem 3 is analogous to Theorem 2. However, we use cross-fitting, which provides an efficient form of data-splitting.

### 3 Conditional Extreme Quantiles

Our analysis thus far focuses on the pseudo-parameters defining the conditional Pareto exponent, which underlines the tail features of the conditional distribution. We now extend our analysis to conditional extreme quantiles.

For  $\tau \in (0, 1)$ , let  $Q_{Y|X, Y > w_n}(\tau)$  denote the conditional  $\tau$ -quantile of  $Y_i$  given  $X_i$  and  $Y_i > w_n$ .

Assumption 1 implies

$$Q_{Y|X, Y > w_n}(\tau) \approx w_n (1 - \tau)^{-\frac{1}{\alpha(X)}} \quad (3.1)$$

with  $\alpha(X) = \exp(X^\top \theta_0)$ . Plug our regularized estimator (2.2) in (3.1) to construct the estimator

$$\hat{Q}_{Y|X, Y > w_n}(\tau) = w_n (1 - \tau)^{-\exp(-X^\top \hat{\theta}}. \quad (3.2)$$

See, for example, Wang, Li, and He (2012) and Wang and Li (2013) for estimators of conditional extreme quantiles under the cases with low- and fixed-dimensional  $X_i$ , and the asymptotic distribution theories therein. Our estimator allows for high dimensionality in  $X_i$ . The following corollary establishes the consistency and convergence rate of our estimator.

**Corollary 1.** Suppose that the Assumptions in Theorem 1 are satisfied. Then, as  $n_0 \rightarrow \infty$ ,

$$\left| \frac{\widehat{Q}_{Y|X,Y>w_n}(\tau)}{Q_{Y|X,Y>w_n}(\tau)} - 1 \right| \lesssim \sqrt{\frac{s_0(\log p)}{n_0}}.$$

In addition, the asymptotic normality can be derived analogously to Theorems 2 and 3. For brevity, we focus on the inference theory for the debiased estimator based on the cross-fitting procedure presented in Section 2.2.2, although the same idea also applies to the debiased estimator based on sample splitting presented in Section 2.2.1. Define  $\tilde{\theta}$  as in  $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_p)^\top$  for  $\tilde{\theta}_j$  defined in (2.10). We can replace  $\widehat{\theta}$  with  $\tilde{\theta}$  to establish the corollary stated below.

Let us introduce the short-handed notation such that

$$q(x^\top \theta) = w_n (1 - \tau)^{-\exp(-x^\top \theta)}$$

for fixed  $\tau$ .

Let  $\widehat{u}_k$  be obtained from the subsample in  $\mathcal{D}_k^c$  using the algorithm (2.7), where in the constraint (2.8),  $e_j$  is replaced with  $x$ . Define the asymptotic variance estimator as follows:

$$\widehat{V}_3 = \frac{1}{K} \sum_{k=1}^K \left[ \widehat{u}_k^\top \left( \frac{1}{n_k} \sum_{i \in \mathcal{D}_k} X_i X_i^\top \right) \widehat{u}_k \right].$$

The following corollary establishes the asymptotic normality for our conditional extreme quantile estimator.

**Corollary 2.** Suppose that Assumptions in Theorem 3 are satisfied. If  $\|x\|_2 \leq C < \infty$ , then

$$\sqrt{n_0} \left\{ \dot{q}(x^\top \widehat{\theta})^2 x^\top \widehat{V}_3 x \right\}^{-1/2} \left\{ q(x^\top \tilde{\theta}) - q(x^\top \theta_0) \right\} \xrightarrow{d} \mathcal{N}(0, 1)$$

as  $n_0, p \rightarrow \infty$ .

Note that the conditional extreme quantile (3.1) is defined conditionally on  $Y_i > w_n$ . Integrating out this condition, (3.1) is equivalent to  $Q_{Y|X}(\tilde{\tau}_n)$ , where

$$\tilde{\tau}_n = 1 - (1 - \tau)(1 - F_{Y|X}(w_n)).$$

By adopting some estimator  $\widehat{F}_{Y|X}(w_n)$ , we can estimate  $Q_{Y|X}(\tilde{\tau}_n)$  accordingly. For instance, we can adapt the conditional density estimators proposed by Efromovich (2010) and Izbicki and Lee (2016, 2017), both of which allow for a high-dimensional covariate vector  $X$ .

## 4 Extension to Large-Scale Online Data

This section extends our proposed HDTIR to online streaming data, addressing computational challenges due to large-scale data. Since online data are collected sequentially, we must update our estimator at each data point, typically using stochastic gradient descent (SGD). Although the X (formerly Twitter) dataset we use in our real data illustration is not large, this extension can be applied to other upcoming research that exploits larger data sets from social media. To highlight sequential data generation, we replace subscript  $i$  with  $t$  in  $\{Y_t, X_t\}$ , which remains i.i.d. across  $t$ .

We modify our HITIR as follows. First, we consider the threshold  $w_n = \bar{w}$  to be predetermined and fixed for the current setting with online streaming data. This threshold can be obtained from a separate sample's empirical quantile of  $Y$ . Otherwise, allowing  $\bar{w}$  to change with data collection makes asymptotic derivation intractable. Second, for tail observations  $Y_t > \bar{w}$ , we effectively have a random sample  $\{Y_t, X_t\}$  from the distribution  $F_{Y,X|Y>\bar{w}}$ . As discussed in Section 2, a sufficiently large  $\bar{w}$  controls the asymptotic bias from deviation from Pareto. To simplify the derivation, we therefore assume

$$1 - F_{Y|Y>\bar{w},X=x}(y) = \left(\frac{y}{\bar{w}}\right)^{-\exp(x^\top \theta_0)}, \quad (4.1)$$

which is an exact Pareto tail above  $\bar{w}$ .

Denote  $T_0 = \sum_{t=1}^T 1 [Y_t > \bar{w}]$  as the effective tail sample size. Considering the tail observations only, we can rewrite the HDTIR problem as

$$\hat{\theta}^{\text{on}} = \arg \min_{\theta} \frac{1}{T_0} \sum_{t=1}^{T_0} \{\exp(X_t^\top \theta) \log(Y_t/\bar{w}) - 1\} + \lambda \|\theta\|_1, \quad (4.2)$$

where the notation  $\hat{\theta}^{\text{on}}$  implies that we are implementing a different algorithm for the online data. Specifically, we propose using the Regularization Annealed epoch Dual AveRaging (RADAR) algorithm (Agarwal et al., 2012), a variant of SGD. Like SGD, RADAR computes the stochastic gradient on one data point at each iteration and provides the optimal convergence rate for the  $L^1$ -norm. Please refer to Agarwal et al. (2012) for more details of the RADAR algorithm.

In addition, we propose to update the debiasing procedure with RADAR as well. Given a fixed  $\bar{w}$ , denote the conditional variance-covariance matrix

$$\Sigma_{\bar{w}} = \mathbb{E}[X_t X_t^\top | Y_t > \bar{w}]$$

and  $\Xi = \Sigma_{\bar{w}}^{-1}$ . Given an estimate  $\widehat{\Xi}$ , we propose the following debiased estimator

$$\tilde{\theta}^{\text{on}} = \widehat{\theta}^{\text{on}} + \frac{1}{T_0} \widehat{\Xi} X^\top Z \left( \widehat{\theta}^{\text{on}} \right),$$

where  $X = [X_1^\top, X_2^\top, \dots, X_{T_0}^\top]^\top$  and  $Z(\theta) = \{Z_1(\theta), \dots, Z_T(\theta)\}^\top$  with

$$Z_t(\theta) = \exp(X_t^\top \theta) \log(Y_t/\bar{w}) - 1.$$

In particular, for the  $j$ -th component,

$$\tilde{\theta}_j^{\text{on}} = \widehat{\theta}_j^{\text{on}} + \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} \left\{ \exp \left( X_t^\top \widehat{\theta} \right) \log(Y_t/\bar{w}) - 1 \right\} X_t.$$

We construct  $\widehat{\Xi}$  by first running the nodewise Lasso

$$\widehat{\gamma}^j = \arg \min_{\gamma^j \in \mathbb{R}^{d-1}} \frac{1}{2T_0} \|X_{\cdot,j} - X_{\cdot,-j}\gamma^j\|_2^2 + \lambda_j \|\gamma^j\|_1, \quad (4.3)$$

where  $X_{\cdot,j}$  is the  $j$ -th column of the matrix  $X$ ,  $X_{\cdot,-j}$  is the design submatrix without the  $j$ -th column, and  $\lambda_j \asymp \sqrt{\log p/T_0}$ . Then, we construct

$$\widehat{\tau}_j = \frac{1}{T_0} (X_{\cdot,j} - X_{\cdot,-j}\widehat{\gamma}^j)^\top X_{\cdot,j}.$$

Given  $\widehat{\gamma}^j$  and  $\widehat{\tau}_j$ , the matrix  $\Xi$  is estimated by

$$\widehat{\Xi} = \widehat{\mathcal{T}} \times \widehat{\mathcal{C}}, \quad (4.4)$$

where  $\widehat{\mathcal{T}} = \text{diag}(1/\widehat{\tau}_1, \dots, 1/\widehat{\tau}_p)$  and

$$\widehat{\mathcal{C}} = \begin{pmatrix} 1 & -\widehat{\gamma}_2^1 & \cdots & -\widehat{\gamma}_p^1 \\ -\widehat{\gamma}_1^1 & 1 & \cdots & -\widehat{\gamma}_p^1 \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\gamma}_1^1 & -\widehat{\gamma}_2^1 & \cdots & 1 \end{pmatrix}.$$

The algorithm to implement the above procedure is provided on the next page.

To study the asymptotic properties, we modify our previous assumptions as follows.

---

**Algorithm** Stochastic optimization based estimation and confidence interval for HDTIR

---

**Inputs:**

Regularization parameter  $\lambda \asymp \sqrt{\log p/T_0}$ , and  $\lambda_j \asymp \sqrt{\log p/T_0}$  for each dimension  $j$ ,  
the noise level  $\sigma$ , confidence level  $1 - \alpha$ , tail threshold  $\bar{w}$

---

**for**  $t = 1$  to  $T$  **do**  
 Randomly sample the data  $(Y_t, X_t)$  and drop this data if  $Y_t < \bar{w}$ .  
 Otherwise, update  $X \leftarrow [X^\top, X_t]^\top$  and  $Y \leftarrow [Y^\top, X_t]^\top$ .  
 Update  $\hat{\theta}$  by running one iteration of RADAR on the optimization problem (4.2)  
 using the stochastic gradient  $\left\{ \exp(X_t^\top \hat{\theta}_t) \log(Y_t/\bar{w}) - 1 \right\} X_t$   
**for**  $j = 1$  to  $p$  **do**  
 Update  $\gamma_t^j$  by running one iteration of RADAR on the optimization problem (4.3)  
 using the stochastic gradient  $(X_{t,-j}^\top \gamma_{t-1}^j - X_{i,j}) X_{t,-j}$   
**end for**

---

**end for**

Let  $\hat{\theta}^{\text{on}}$  and  $\hat{\gamma}^j$  for  $j \in \{1, \dots, d\}$  be the final outputs.

Construct the debiased estimator  $\tilde{\theta}^{\text{on}}$  with  $\hat{\Xi}$  defined in (4.4)

$$\tilde{\theta}^{\text{on}} = \hat{\theta}^{\text{on}} + \frac{1}{T_0} \hat{\Xi} X^\top Z(\hat{\theta}^{\text{on}})$$


---

**Outputs:**

The estimator  $\tilde{\theta}^{\text{on}}$  and the  $(1 - \alpha)$  confidence interval for each  $\theta_j^*$ :

$$\tilde{\theta}_j^{\text{on}} \pm z_{\alpha/2} \sqrt{(\hat{\Xi} \hat{\Sigma} \hat{\Xi}^\top)_{j,j} / T_0}, \text{ where } \hat{\Sigma} = \frac{1}{T_0} X^\top X$$


---

**Assumption 5** (Online).  $\{Y_t, X_t\}$  is i.i.d. with distribution satisfying (4.1). For all  $j$ ,  $X_{t,j}$  has a compact support  $\mathcal{X}_j$  with  $\sup_{x \in \mathcal{X}_j} f_{X_{t,j}|Y_t>\bar{w}}(x) < \bar{f} < 0$ . For some constants  $C_1 > 0, C_2 > 1$ , the parameter space satisfies

$$\Omega(s_0) = \left\{ \begin{array}{l} (\theta, \Sigma_{\bar{w}}) : \|\theta\| \leq s_0, \|\theta\|_1 \leq C_1, \\ 1 < \underline{\alpha} \leq \inf_x \exp(x^\top \theta) \leq \sup_x \exp(x^\top \theta) \leq \bar{\alpha} < \infty, \\ C_2^{-1} \leq \lambda_{\min}(\Sigma_{\bar{w}}) \leq \lambda_{\max}(\Sigma_{\bar{w}}) \leq C_2. \end{array} \right\}.$$

**Theorem 4.** Suppose Assumption 4 holds and  $s_0 \ll \sqrt{T_0} / (\sqrt{\log T_0} \log p)$ . Using the same algorithm parameters as in Proposition 1 in Agarwal, Negahban and Wainwright (2012), it holds that for all  $j = 1, \dots, p$ ,

$$\frac{\sqrt{T_0} (\tilde{\theta}_j^{\text{on}} - \theta_{0j})}{\sqrt{(\hat{\Xi}^\top \hat{\Sigma} \hat{\Xi})_{j,j}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

## 5 Simulation Studies

In this section, we use simulated data to numerically evaluate the performance of our proposed method of estimation and inference. Two designs for the  $p$ -dimensional parameter vector  $\theta_0$  are employed:

1. Sparse Design:  $\theta_0 = (1.0, 0.9, 0.8, \dots, 0.2, 0.1, 0.0, 0.0, 0.0, \dots)^\top$ , and
2. Exponential Design:  $\theta_0 = (1.0, 0.5, 0.5^2, 0.5^3, \dots)^\top$ .

A random sample of  $(Y_i, X_i^\top)^\top$  is generated as follows. Three designs for the  $p$ -dimensional covariate vector  $X_i$  are employed:

1. Gaussian Design:  $X_i \sim N(0, 0.1^2 \cdot I_p)$ ,
2. Uniform Design:  $X_i \sim \text{Uniform}(-0.1, 0.1)$  and
3. Bernoulli Design:  $X_i \sim 0.1 \cdot \text{Bernoulli}(0.1)$ .

where  $I_p$  denotes the  $p \times p$  identity matrix. In turn, generate the exponents by

$$\alpha_i = \exp(X_i^\top \theta_0)$$

and then generate  $Y_i$  by

$$Y_i = \Lambda^{-1}(U_i; \alpha_i), \quad U_i \sim \text{Uniform}(0, 1),$$

where  $\Lambda(\cdot; \alpha)$  denotes the CDF of the Pareto distribution with the unit scale and exponent  $\alpha$ .

In each iteration, we draw a random sample  $(Y_i, X_i^\top)^\top$  of size  $n = 10,000$ . Setting the cutoff  $\omega$  to the 95-th empirical percentile of  $\{Y_i\}_{i=1}^n$  we have the effective sample size of  $n_0 = 500$  from five percent of  $N$ . We vary the dimension  $p \in \{250, 500, 1000\}$  of the parameter vector  $\theta_0$  across sets of simulations. While there are  $p$  coordinates in  $\theta_0$ , we focus on the first coordinate  $\theta_{01} = 1.0$  for evaluating our method of estimation and inference. Throughout, we use  $K = 5$  for the number of subsamples in sample splitting. The other tuning parameters are set according to Assumption 3 where  $c = 1$ ,  $c' = 1$  and  $c'' = 100$ . We run 10,000 Monte Carlo iterations for each design.

Table 1 summarizes the simulation results. The sets of results vary with the effective sample size  $n_0$ , the dimension  $p$  of the parameter vector  $\theta$ , the design for the parameter vector  $\theta$ , and the design for the covariate vector  $X$ . For each row, displayed are the bias (Bias) of the debiased estimator  $\tilde{\theta}_1$ ,

$n_0$	$p$	$\theta$	$X$	$\Lambda$	Bias	SD	RMSE	95%
500	250	Sparse	Gaussian	Pareto	0.023	0.545	0.545	0.918
500	250	Exponential	Gaussian	Pareto	0.012	0.494	0.494	0.936
500	500	Sparse	Gaussian	Pareto	0.022	0.542	0.542	0.919
500	500	Exponential	Gaussian	Pareto	0.007	0.487	0.487	0.942
500	1000	Sparse	Gaussian	Pareto	0.033	0.520	0.521	0.927
500	1000	Exponential	Gaussian	Pareto	0.000	0.483	0.483	0.938
500	250	Sparse	Uniform	Pareto	0.000	0.824	0.824	0.941
500	250	Exponential	Uniform	Pareto	-0.018	0.799	0.799	0.949
500	500	Sparse	Uniform	Pareto	-0.004	0.825	0.825	0.941
500	500	Exponential	Uniform	Pareto	-0.015	0.791	0.792	0.948
500	1000	Sparse	Uniform	Pareto	-0.008	0.820	0.820	0.941
500	1000	Exponential	Uniform	Pareto	-0.020	0.800	0.800	0.944
500	250	Sparse	Bernoulli	Pareto	-0.233	0.716	0.753	0.961
500	250	Exponential	Bernoulli	Pareto	-0.101	0.831	0.837	0.955
500	500	Sparse	Bernoulli	Pareto	-0.252	0.710	0.753	0.964
500	500	Exponential	Bernoulli	Pareto	-0.112	0.823	0.830	0.958
500	1000	Sparse	Bernoulli	Pareto	-0.244	0.713	0.753	0.963
500	1000	Exponential	Bernoulli	Pareto	-0.101	0.826	0.832	0.955

Table 1: Simulation results. The sets of results vary with the dimension  $p$  of the parameter vector  $\theta_0$ , the design for the parameter vector  $\theta_0$ , and the design for the covariate vector  $X$ . For each row, displayed are the bias (Bias), standard deviations (SD), root mean square errors (RMSE), and the coverage frequencies by the 95% confidence interval (95%).

standard deviations (SD), root mean square errors (RMSE), and the coverage frequencies by the 95% confidence interval (95%).

For each set, the bias is much smaller than the standard deviations and hence the 95% confidence interval delivers accurate coverage frequencies. We ran many other sets of simulations with different values of  $n_0$  and  $p$  as well as parameter designs and data generating designs, and confirm that the simulation results turned out to be similar in qualitative patterns to those presented here. The additional results are omitted from the paper to avoid repetitive exposition.

## 6 Application: Text Analysis of Viral Posts about LGBTQ+

In this section, we apply our proposed method to analyze LGBTQ+-related posts on X (formerly Twitter). Our goal is to infer the impact of specific words on attracting “likes” for these posts. The dataset comprises tweets containing the keyword “LGBT,” scraped from Twitter between August 21 and August 26, 2022.<sup>1</sup>

Each observation in our study represents a single post. Our sample includes a total of  $n = 32,456$  posts. The data records the number of likes,  $Y_i$ , that the  $i$ -th post has received. As we will demonstrate below,  $Y_i$  follows a heavy-tailed distribution: most posts attract a small number of likes, while a few viral posts garner a large number of likes. We constructed a word bank consisting of 936,556 unique words used across the  $n = 32,456$  posts in our sample. The  $j$ -th coordinate,  $X_{ij}$ , of the covariate vector  $X_i$  takes a value of 1 if the  $j$ -th word in the word bank is used in the  $i$ -th post and 0 otherwise. From these 936,556 unique words, we only include the 500 most frequently used words to create the binary indicators in  $X_i$ . Therefore, the dimension  $p$  of  $X_i$  is 500. This list excludes explicitly articles, auxiliary verbs, and prepositions.

Figure 1 in the introduction presents the log-log plot of the empirical distribution  $Y_{i=1}^n$  for posts with a positive number of likes. We focus on posts with positive likes because the logarithm of zero is undefined. The horizontal axis represents the rank of  $Y$ , while the vertical axis represents  $\log(Y)$ . The approximate linearity of this log-log plot suggests that the distribution of  $Y$  follows a power law, indicating that  $Y$  is characterized by a Pareto distribution.

Table 2 displays the 30 most frequently used words in LGBTQ+ posts. For each word, the total number of times it appeared (Count) and the total number of posts in which it appeared (Tweets) are shown. The last value corresponds to  $\sum_{i=1}^n X_{ij}$ . All characters have been converted to lowercase to ensure the counting is not case-sensitive. Notice that the most frequent word, “lgbt,” is distinct from the eleventh most frequent word, “#lgbt.” The former is a plain word, while the latter functions as a hashtag, serving to link posts with others containing the same hashtag.

We apply our proposed method of estimation and inference to analyze the effects of using these and other words on the tail shape of the distribution of the number of likes. Consistent with our simulation

---

<sup>1</sup>The data set is publicly available at <https://www.kaggle.com/datasets/vencerlanz09/lgbt-tweets>.

	Word	Count	Tweets		Word	Count	Tweets		Word	Count	Tweets
1	lgbt	23734	8133	11	#lgbt	4989	1280	21	it's	3194	1842
2	and	18706	12357	12	this	4956	3532	22	lgbt+	3165	578
3	i	10162	1559	13	community	4189	3640	23	their	3127	2641
4	that	9022	6931	14	have	4186	3605	24	my	2747	2003
5	you	8965	5374	15	just	3563	2900	25	don't	2690	2195
6	people	7117	5495	16	or	3555	2903	26	what	2662	1840
7	it	7015	5025	17	so	3517	2534	27	he	2631	1548
8	not	5987	4650	18	if	3346	2207	28	your	2608	2068
9	they	5695	3705	19	all	3231	2624	29	gay	2559	1833
10	but	5012	3770	20	who	3227	2723	30	do	2559	2072

Table 2: The top 30 most frequent words used in LGBTQ+ posts. Displayed next to each word are the total number of times it appeared (Count) and the total number of posts in which it appeared (Tweets). All the characters are unified to lower-case letters so that the counting is not case sensitive.

studies, we set  $w_n$  to the 95th percentile of the empirical distribution of  $Y_{i=1}^n$ , which results in an effective sample size of  $n_0 = 1,623$ . The rules for selecting the tuning parameters remain the same as those used in our simulation studies.

Table 3 presents the estimates, standard errors, 95% confidence intervals, and t-statistics for  $\theta_j$  for the 30 most frequently used words, listed in the same order as in Table 2. Notably, the most frequent word, “lgbt,” has a significantly negative coefficient, while the eleventh most frequent word, “#lgbt,” has a significantly positive coefficient. Recall that smaller values of the Pareto exponent correspond to more extreme values of  $Y_i$ . Therefore, this finding suggests that using the plain word “lgbt” tends to attract a substantially larger number of likes, whereas using the hashtag “#lgbt” may have the opposite effect. Most of the other words in Table 3 are statistically insignificant, with the exceptions of “they” and “it's,” whose positive coefficients indicate their adverse effects.

We then identify the 10 most effective words and the 10 least effective words from the list of  $p = 500$  words. Table 4 presents the estimates, standard errors, 95% confidence intervals, and t-statistics for  $\theta_j$  for these words. The words are sorted in descending order based on the absolute value of the estimate  $\tilde{\theta}_j$ . Again, we find that the plain word “lgbt” is the only significantly effective word. In contrast, hashtags containing this effective keyword, such as “#lgbtqia” and “#lgbtq,” tend to have negative

$j$	Word	$\tilde{\theta}_j$	SE	95% CI		t	$j$	Word	$\tilde{\theta}_j$	SE	95% CI		t
1	lgbt	-0.14	0.06	[-0.26	-0.03]	-2.40	16	or	0.00	0.09	[-0.17	0.18]	0.04
2	and	-0.01	0.05	[-0.11	0.10]	-0.16	17	so	0.10	0.10	[-0.09	0.29]	1.02
3	i	0.07	0.12	[-0.17	0.31]	0.58	18	if	-0.08	0.11	[-0.30	0.14]	-0.73
4	that	0.04	0.06	[-0.08	0.16]	0.70	19	all	0.17	0.09	[0.00	0.35]	1.96
5	you	0.10	0.07	[-0.03	0.24]	1.49	20	who	0.10	0.08	[-0.05	0.25]	1.26
6	people	0.07	0.07	[-0.06	0.20]	1.04	21	it's	0.22	0.10	[0.02	0.42]	2.14
7	it	0.06	0.07	[-0.08	0.21]	0.89	22	lgbt+	0.03	0.17	[-0.31	0.37]	0.18
8	not	0.01	0.08	[-0.14	0.16]	0.10	23	their	0.06	0.09	[-0.12	0.24]	0.68
9	they	0.16	0.07	[0.02	0.30]	2.26	24	my	0.17	0.10	[-0.02	0.37]	1.75
10	but	0.02	0.08	[-0.14	0.18]	0.25	25	don't	0.14	0.10	[-0.05	0.34]	1.43
11	#lgbt	0.41	0.15	[0.11	0.71]	2.69	26	what	0.09	0.11	[-0.13	0.31]	0.81
12	this	0.13	0.07	[-0.01	0.28]	1.82	27	he	0.04	0.09	[-0.15	0.22]	0.41
13	community	-0.03	0.08	[-0.19	0.13]	-0.38	28	your	0.07	0.10	[-0.13	0.27]	0.71
14	have	0.12	0.08	[-0.03	0.27]	1.60	29	gay	0.08	0.10	[-0.11	0.27]	0.82
15	just	-0.04	0.09	[-0.22	0.15]	-0.41	30	do	0.11	0.10	[-0.09	0.32]	1.09

Table 3: Estimates, standard errors, 95% confidence intervals, and the t statistics for  $\theta_j$  for the top 30 most frequent words. These words are listed in the same order as in Table 2.

contributions to attracting likes.

## 7 Summary

This paper introduces a novel high-dimensional tail index regression (HDTIR) model inspired by observing power-law distributions in social media posts, particularly in the distribution of “likes” on viral content. We tackle the challenges of estimating and inferring the parameters of the tail index model when the dimension of the explanatory variables increases and may exceed the sample size.

We begin by developing a regularized estimation method for the HDTIR model, demonstrating its consistency and establishing its convergence rate. To facilitate inference, we introduce a debiasing technique that corrects the bias introduced by regularization. This allows us to derive the asymptotic normality of the debiased estimator, providing a robust framework for statistical inference in high-dimensional settings.

The methodology is further extended to accommodate large-scale online data, particularly relevant

10 Most Effective Words					10 Least Effective Words						
Word	$\tilde{\theta}_j$	SE	95% CI		t	Word	$\tilde{\theta}_j$	SE	95% CI		t
lgbt	-0.14	0.06	[-0.26	-0.03]	-2.40	lgb	4.04	0.80	[2.48	5.61]	5.06
if	-0.08	0.11	[-0.30	0.14]	-0.73	ukraine	3.68	0.67	[2.36	5.00]	5.46
me	-0.07	0.11	[-0.28	0.14]	-0.67	377a	3.30	0.70	[1.92	4.68]	4.69
make	-0.07	0.14	[-0.33	0.20]	-0.49	#lgbtqia	3.01	0.69	[1.67	4.36]	4.39
just	-0.04	0.09	[-0.22	0.15]	-0.41	#pride	2.74	0.64	[1.48	3.99]	4.27
community	-0.03	0.08	[-0.19	0.13]	-0.38	let's	2.62	0.73	[1.18	4.06]	3.57
and	-0.01	0.05	[-0.11	0.10]	-0.16	#lgbtq	2.60	0.54	[1.53	3.66]	4.79
also	0.00	0.14	[-0.28	0.27]	-0.01	american	2.42	0.54	[1.36	3.49]	4.46
has	0.00	0.10	[-0.19	0.19]	-0.01	magic	2.41	0.55	[1.34	3.49]	4.41
or	0.00	0.09	[-0.17	0.18]	0.04	x	2.33	0.45	[1.44	3.21]	5.16

Table 4: Estimates, standard errors, 95% confidence intervals, and the t statistics for  $\theta_j$  for the top 30 most effective words to attract likes. These words are sorted in descending order in terms of the estimate  $\tilde{\theta}_j$ .

in the context of social media, where data streams are large and continuously generated. We employ a variation of stochastic gradient descent to efficiently handle these large datasets, ensuring that the proposed methods can scale to meet the demands of real-world applications.

Extensive simulation studies validate the theoretical properties of our model, showing strong performance even in finite samples. Finally, we apply the HDTIR method to a dataset of viral posts on X (formerly Twitter) related to LGBTQ+ topics. This empirical analysis reveals insights into how specific words influence the likelihood of a post going viral, with terms like ‘lgbt’ playing a significant role while hashtags like ‘#lgbtq’ do not. The results demonstrate the practical utility of the HDTIR model in understanding and predicting the factors that drive the popularity of online content.

# Appendix

This appendix collects mathematical proofs and technical details. Appendix A presents the proofs of main theorems. Appendix B presents some technical lemmas. Appendix C presents the proofs of the corollaries.

## A Proofs for the Main Theorems

### A.1 Proof of Theorem 1

*Proof of Theorem 1.* We define three events:

$$\begin{aligned}\mathcal{E}_1 &= \left\{ \|\hat{\theta} - \theta_0\|_2 \lesssim \sqrt{\frac{s_0(\log p)}{n}} \right\}, \\ \mathcal{E}_2 &= \left\{ \|\hat{\theta} - \theta_0\|_1 \lesssim \sqrt{\frac{s_0^2(\log p)}{n_0}} \right\}, \text{ and} \\ \mathcal{E}_3 &= \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \left[ X_i^\top (\hat{\theta} - \theta_0) \right]^2 \lesssim \frac{s_0(\log p)}{n_0} \right\}.\end{aligned}$$

We first show  $\mathbb{P}(\mathcal{E}_1) \geq 1 - p^{-c}$  for some constant  $c$ . To this end, we resort to Proposition 1 on properties of the Lasso. Without loss of generality, we order the data such that  $Y_i > w_n$  for  $i = 1, \dots, n_0$ . Then we can write

$$\begin{aligned}\dot{\ell}_{n_0}(\theta_0) &= \frac{1}{n_0} \sum_{i=1}^n \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i \times 1 \{Y_i > w_n\} \\ &= \frac{1}{n_0} \sum_{i=1}^{n_0} \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i.\end{aligned}$$

Define

$$Z_{n_0,i,j} = \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_{i,j}.$$

Note that  $\{Z_{n_0,i,j}\}_{i=1}^{n_0}$  are still i.i.d. conditioning on  $Y_i > w_n$  for  $i = 1, \dots, n_0$ . We keep the subscript  $n_0$  to emphasize that we consider the subsample  $\{Y_i > w_n\}_{i=1}^{n_0}$ .

We are now going to verify (i)  $\|\dot{\ell}_{n_0}(\theta_0)\|_\infty \lesssim \sqrt{(\log p)/n_0}$  with probability at least  $1 - p^{-c}$ ; and (ii) for  $F(\varsigma, S; \psi, \psi_0)$  defined in Proposition 1,  $F(\varsigma, S; \psi, \psi_0) \gtrsim s_0^{-1/2}$  with probability at least  $1 - p^{-c}$ .

For (i), Lemma 1 below and Vershynin (2010, Remark 5.18) imply that  $Z_{n_0,i,j} - \mathbb{E}[Z_{n_0,i,j}]$  is subexponential. Therefore, by concentration inequality for subexponential random variables (see, e.g., Proposition 5.16 of Vershynin (2010)) and the union bound,

$$\begin{aligned} & \mathbb{P} \left( \max_{1 \leq j \leq p} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \{Z_{n_0,i,j} - \mathbb{E}[Z_{n_0,i,j}]\} \right| \geq t \middle| \{Y_i > w_n\}_{i=1}^{n_0} \right) \\ & \lesssim p \exp \left( -n_0 \min \left\{ \frac{t^2}{2}, \frac{t}{2} \right\} \right). \end{aligned}$$

For  $t \leq 1$  and hence  $\frac{t^2}{2} < \frac{t}{2}$ , set

$$\delta = \exp \left( \log p - \frac{n_0 t^2}{2} \right),$$

implying that  $t = \sqrt{\frac{2 \log(p/\delta)}{n_0}}$ . When  $p > n_0$ , set  $\delta = p^{-c}$  and we have  $t = \sqrt{\frac{2 \log(p^{1+c})}{n_0}} \sim \sqrt{\frac{\log p}{n_0}}$ . Similarly, when  $p < n_0$ , set  $\delta = n_0^{-c}$ , we have  $t = v \sqrt{\frac{2 \log p + c \log n_0}{n_0}} \sim \sqrt{\frac{\log n_0}{n_0}}$ . Therefore,

$$\mathbb{P} \left( \max_{1 \leq j \leq p} \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \{Z_{n_0,i,j} - \mathbb{E}[Z_{n_0,i,j}]\} \right| \geq C \sqrt{\frac{(\log p)}{n_0}} \middle| \{Y_i > w_n\}_{i=1}^{n_0} \right) \leq p^{-c}$$

for some  $c, C \in \mathbb{R}^+$ . Moreover, Lemma 3.(ii) and Assumption 3 imply that

$$|\mathbb{E}[Z_{n_0,i,j}|Y_i > w_n]| \leq C w_n^{-\beta} = o \left( \sqrt{\frac{(\log p)}{n_0}} \right).$$

Therefore, (i) holds with probability at least  $1 - p^{-c}$  conditional on  $\{Y_i > w_n\}_{i=1}^{n_0}$  and hence also unconditionally.

We are now going to verify (ii). Because

$$\ddot{\ell}_{n_0}(\theta_0) = \frac{1}{n_0} \sum_{i=1}^{n_0} \exp(X_i^\top \theta_0) \log(Y_i/w_n) X_i X_i^\top$$

for  $0 < \eta < 1$ ,

$$\begin{aligned} \langle b, \dot{\ell}_{n_0}(\theta_0 + b) - \dot{\ell}_{n_0}(\theta_0) \rangle &= \int_0^1 \langle b, \ddot{\ell}_{n_0}(\theta_0 + tb) \rangle dt \\ &= \int_0^1 \frac{1}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_i) \exp(X_i^\top (\theta_0 + tb)) (b^\top X_i)^2 dt. \end{aligned}$$

Let  $\psi(b) = \psi_0(b) = \|b\|_2$ . For  $F(\varsigma, S; \psi, \psi_0)$  defined in Proposition 1 and  $S = \{j : \theta_{0j} \neq 0\}$ , for some

constant  $M > 0$ ,

$$\begin{aligned}
F(\varsigma, S; \psi, \psi_0) &= \inf_{b \in \mathcal{C}(\varsigma, S), \psi_0(b) \leq 1} \frac{\left\langle b, \dot{\ell}_{n_0}(\theta_0 + b) - \dot{\ell}_{n_0}(\theta_0) \right\rangle \exp\left(-\psi_0(b)^2 - \psi_0(b)\right)}{\|b_S\|_1 \|b\|_2} \\
&= \inf_{b \in \mathcal{C}(\varsigma, S), \psi_0(b) \leq 1} \frac{\int_0^1 \frac{1}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_i) \exp(X_i^\top (\theta_0 + tb)) (b^\top X_i)^2 dt \exp\left(-\|b\|_2^2 - \|b\|_2\right)}{\|b_S\|_1 \|b\|_2} \\
&\geq M \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2 \leq 1} \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\int_0^1 \log(Y_i/w_i) \exp(X_i^\top (\theta_0 + tb)) (b^\top X_i)^2 dt}{\|b_S\|_1 \|b\|_2} \\
&= M \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2 = 1} \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\log(Y_i/w_i) \alpha(X_i) [\exp(X_i^\top b) - 1] (b^\top X_i)^2}{\|b_S\|_1 \|b\|_2 (b_i^\top X)} \\
&\geq s_0^{-1/2} M \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2 = 1} \frac{1}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_i) (b^\top X_i)^2,
\end{aligned}$$

where the last inequality follows by  $\alpha(X_i) > \underline{\alpha} > 0$  and the fact that  $\exp(x) - 1 \geq x$  for any  $x \in \mathbb{R}$ .

To proceed, we define the truncation function  $\varphi_T(\cdot)$  for some constant  $T > 1$  such that for any  $x > 0$

$$\varphi_T[x] = \begin{cases} x & \text{if } x \leq T \\ 2T - x & \text{if } x > T. \end{cases}$$

Then, we also have

$$F(\varsigma, S; \psi, \psi_0) \geq s_0^{-1/2} M \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2 = 1} \frac{1}{n_0} \sum_{i=1}^{n_0} \varphi_T[\log(Y_i/w_i)] \varphi_T[(b^\top X_i)^2].$$

Lemma 4 below shows that, for some constants  $c, T > 0$ ,

$$\mathbb{P}\left(\inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2 = 1} \frac{1}{n_0} \sum_{i=1}^{n_0} \varphi_T[\log(Y_i/w_i)] \varphi_T[(b^\top X_i)^2] > c \middle| \{Y_i > w_n\}_{i=1}^{n_0}\right) \geq 1 - e^{-cn_0 - \log p},$$

which implies (ii).

Thus, by (i), (ii), and the fact that the negative log-likelihood  $\ell_{n_0}(\theta)$  is a convex function, Proposition 1 holds with  $z^* = \|\dot{\ell}_{n_0}(\theta_0)\|_\infty$  and  $\lambda_{n_0} \asymp \sqrt{(\log p)/n_0}$ , yielding that  $\mathcal{E}_1$  holds with probability at least  $1 - p^{-c}$ .

For  $\mathcal{E}_2$ , since all conditions in Proposition 1 are satisfied, Lemma 7 in Cai et al. (2023) implies that

$$\|\widehat{\theta} - \theta_0\|_1 \leq (1 + \varsigma) \left\| \left( \widehat{\theta} - \theta_0 \right)_S \right\|_1 \leq (1 + \varsigma) \sqrt{s_0} \left\| \left( \widehat{\theta} - \theta_0 \right)_S \right\|_2 \lesssim \sqrt{\frac{s_0 (\log p)}{n_0}}$$

with probability at least  $1 - p^{-c}$ . Then,  $\mathcal{E}_2$  holds with probability at least  $1 - p^{-c}$ .

For  $\mathcal{E}_3$ , by Assumption 2,

$$\begin{aligned} \frac{1}{n} \left\| X (\widehat{\theta} - \theta_0) \right\|_2 &= \frac{1}{n} (\widehat{\theta} - \theta_0)^\top X^\top X (\widehat{\theta} - \theta_0) \\ &\lesssim \left\| \widehat{\theta} - \theta_0 \right\|_2^2 \\ &\lesssim \frac{s_0 (\log p)}{n_0} \end{aligned}$$

with probability at least  $1 - p^{-c}$ . Thus,  $\mathcal{E}_3$  holds with probability at least  $1 - p^{-c}$ .  $\square$

Below, we cite the auxiliary proposition from the existing literature, which we use to prove our first main theorem.

**Proposition 1** (Huang and Zhang, 2012 and Cai et al., 2023). *Let  $\widehat{\theta} = \arg \min_{\theta} \{\ell_n(\theta) + \lambda \|\theta\|_1\}$  be the Lasso estimator for some generalized linear model with true regression coefficient  $\theta_0$ , where the normalized negative log-likelihood  $\ell(\theta)$  is a convex function. Let*

$$F(\varsigma, S; \psi, \psi_0) = \inf_{b \in \mathcal{C}(\varsigma, S), \psi_0(b) \leq 1} \frac{b^\top (\dot{\ell}_n(\theta_0 + b) - \dot{\ell}_n(\theta_0)) e^{-\psi_0^2(b) - \psi_0(b)}}{\|b_S\|_1 \psi(b)},$$

where  $S = \{j : \theta_{0j} \neq 0\}$ ,  $\psi$  and  $\psi_0$  are semi-norms,  $M_2 > 0$  is a constant, and

$$\mathcal{C}(\varsigma, S) = \{b \in \mathbb{R}^p : \|b_{S^c}\|_1 \leq \varsigma \|b_S\|_1 \neq 0\}.$$

Define

$$\Omega = \left\{ \frac{\lambda + z^*}{(\lambda - z^*)_+} \leq \xi, \frac{\lambda + z^*}{F(\varsigma, S; \psi, \psi_0)} \leq \eta e^{-\eta^2 - \eta} \right\}$$

for some  $\eta \leq 1/2$  and  $z^* = \|\dot{\ell}_{n_0}(\theta_0)\|_\infty \leq \lambda$ . Then, in the event  $\Omega$ , we have  $\psi(\widehat{\theta} - \theta_0) \leq \frac{(\lambda + z^*)e^{\eta^2 + \eta}}{F(\varsigma, S; \psi, \psi_0)}$ .

## A.2 Proof of Theorem 2

*Proof of Theorem 2.* The mean value expansion yields that, for each  $j = 1, \dots, p$ ,

$$\begin{aligned} \widetilde{\theta}_j - \theta_{0j} &= \widehat{\theta}_j - \theta_{0j} - \frac{\widehat{u}_j^\top}{n_0} \sum_{i=1}^{n_0} \left\{ \exp(X_i^\top \widehat{\theta}) \log(Y_i/w_n) - 1 \right\} X_i \\ &= -\frac{\widehat{u}_j^\top}{n_0} \sum_{i=1}^{n_0} \left\{ \exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1 \right\} X_i \\ &\quad - \left( \frac{\widehat{u}_j^\top}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_n) \{\exp(X_i^\top \theta_0)\} X_i X_i^\top - e_j \right) (\widehat{\theta} - \theta_0) \\ &\quad + \frac{\widehat{u}_j^\top}{n_0} \sum_{i=1}^{n_0} \log(Y_i/w_n) X_i \left\{ \exp \left( X_i^\top \widehat{\theta} + t X_i^\top (\theta_0 - \widehat{\theta}) \right) \right\} \left[ X_i^\top (\widehat{\theta} - \theta_0) \right]^2, \end{aligned}$$

for some  $0 < t < 1$ .

Let  $W_i = \widehat{u}_j^\top \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i$ . Note that, conditional on  $\{X_i\}_{i \in \mathcal{D}_1}$ ,  $\widehat{u}_j$  is non-stochastic. Then, Lemma 3.(i) yields that

$$\begin{aligned} & \mathbb{E}[W_i | \{X_i\}_{i \in \mathcal{D}_1}, Y_i > w_n] \\ &= \mathbb{E}\left[\widehat{u}_j^\top \{\exp(X_i^\top \theta_0) \mathbb{E}[\log(Y_i/w_n) | X_i, Y_i > w_n] - 1\} X_i | \{X_i\}_{i \in \mathcal{D}_1}, Y_i > w_n\right] \\ &= \mathbb{E}\left[\widehat{u}_j^\top X_i \times O\left(w_n^{-\beta(X_i)}\right) | \{X_i\}_{i \in \mathcal{D}_1}, Y_i > w_n\right] \\ &\leq \mathbb{E}\left[\widehat{u}_j^\top X_i | \{X_i\}_{i \in \mathcal{D}_1}, Y_i > w_n\right] O\left(w_n^{-\frac{\beta}{2}}\right) \\ &= O\left(w_n^{-\frac{\beta}{2}}\right). \end{aligned}$$

Similarly, Lemma 3.(iii) yields that

$$\begin{aligned} & \mathbb{E}[W_i^2 | \{X_i\}_{i \in \mathcal{D}_1}, X_i, Y_i > w_n] \\ &= \left(\widehat{u}_j^\top X_i\right)^2 \left(1 + O\left(w_n^{-\frac{\beta}{2}}\right)\right). \end{aligned}$$

The rest follows from a similar argument to the proof of Theorem 3.  $\square$

### A.3 Proof of Theorem 3

*Proof of Theorem 3.* The mean value expansion yields that, for each  $j = 1, \dots, p$ ,

$$\begin{aligned} & \widetilde{\theta}_j - \theta_{0j} \\ &= \frac{1}{K} \sum_{k=1}^K (\widetilde{\theta}_{j,k} - \theta_{0j}) \\ &= \frac{1}{K} \sum_{k=1}^K \left\{ \widehat{\theta}_{j,k} - \theta_{0j} - \frac{\widehat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \left\{ \exp(X_i^\top \widehat{\theta}_k) \log(Y_i/w_n) - 1 \right\} X_i \right\} \\ &= -\frac{1}{K} \sum_{k=1}^K \frac{\widehat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i \\ &\quad - \frac{1}{K} \sum_{k=1}^K \left( \frac{\widehat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \exp(X_i^\top \theta_0) \log(Y_i/w_n) X_i X_i^\top - e_j \right) (\widehat{\theta}_k - \theta_0) \\ &\quad + \frac{1}{K} \sum_{k=1}^K \left( \frac{\widehat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \log(Y_i/w_n) X_i \exp\left(X_i^\top \widehat{\theta}_k + t X_i^\top (\theta_0 - \widehat{\theta}_k)\right) \cdot [X_i^\top (\widehat{\theta}_k - \theta_0)]^2 \right) \\ &\equiv -I_1 - I_2 + I_3 \end{aligned}$$

for some  $t \in (0, 1)$ . We first show that  $\sqrt{n_0}I_2 = o_p(1)$  and  $\sqrt{n_0}I_3 = o_p(1)$ .

For  $I_2$ , note that  $n_0 = Kn_k$  and

$$\begin{aligned} I_2 &\leq \left\| \frac{1}{K} \sum_{k=1}^K \left( \frac{\hat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \exp(X_i^\top \theta_0) \log(Y_i/w_n) X_i X_i^\top - e_j \right) \right\|_\infty \max_{1 \leq k \leq K} \|\hat{\theta}_k - \theta_0\|_1 \\ &= O_p \left( \gamma_{1n_0} \cdot \max_{1 \leq k \leq K} \|\hat{\theta}_k - \theta_0\|_1 \right) = O_p \left( \frac{s_0 \log p}{n_0} \right), \end{aligned}$$

where the first equality follows from Lemma 2, so we have  $\sqrt{n_0}I_2 = o_p(1)$  from the condition in the theorem.

For  $I_3$ , define

$$\Delta_i = \log(Y_i/w_n) \exp \left( X_i^\top \hat{\theta}_k + t X_i^\top (\theta_0 - \hat{\theta}_k) \right) \cdot \left[ X_i^\top (\hat{\theta}_k - \theta_0) \right]^2.$$

By Cauchy-Schwartz inequality, for each  $j = 1, \dots, p$ ,

$$\begin{aligned} \left| \frac{1}{\sqrt{n_0}} \sum_{k=1}^K \hat{u}_{j,k}^\top \sum_{i=1}^{n_k} X_i \Delta_i \right| &\leq \max_{1 \leq i \leq n_k} |\hat{u}_{j,k}^\top X_i| \cdot \left| \frac{1}{\sqrt{n_0}} \sum_{k=1}^K \sum_{i=1}^{n_k} \Delta_i \right| \\ &\lesssim C \frac{\gamma_{2n_0}}{\sqrt{n_0}} \sum_{k=1}^K \sum_{i=1}^{n_k} \left[ X_i^\top (\hat{\theta}_k - \theta_0) \right]^2 (1 + o_p(1)) \\ &\lesssim C \frac{\gamma_{2n_0} s_0 \log p}{\sqrt{n_0}} \\ &= O_p \left( \frac{s_0 \log p \sqrt{\log n_0}}{\sqrt{n_0}} \right) = o_p(1), \end{aligned}$$

where the second inequality follows from the constraint  $\max_{1 \leq i \leq n_k} |\hat{u}_{j,k}^\top X_i| \leq \gamma_{2n_0}$  and Lemma 5, the third inequality follows from Theorem 1, the first equality follows because of Assumption 3(iii), and the second equality follows from the condition in the theorem. Hence,  $\sqrt{n_0}I_3 = o_p(1)$ .

Next, we derive the asymptotic normality result for  $I_1$ . Define

$$v_j = \frac{1}{K} \sum_{j=1}^K \hat{u}_{j,k}^\top \left[ \frac{1}{n_k} \sum_{i=1}^{n_k} X_i X_i^\top \right] \hat{u}_{j,k}^\top$$

and

$$\Psi_{i,k} = v_j^{-1/2} \hat{u}_{j,k}^\top \{ \exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1 \} X_i.$$

Conditional on  $\{X_i\}_{i \in \mathcal{D}_k}$ ,  $\hat{u}_{j,k}$  is non-stochastic. Also conditional on  $X_i$  and  $Y_i > w_n$ ,  $\{\Psi_{i,k}\}_{i \in \mathcal{D}_k}$  are

i.i.d. with

$$\begin{aligned}\mathbb{E} [\Psi_{i,k} | \{X_i\}_{i \in \mathcal{D}_k}, Y_i > w_n] &= O\left(w_n^{-\beta}\right) \\ \sum_{i \in \mathcal{D}_k} \mathbb{E} [\Psi_{i,k}^2 | \{X_i\}_{i \in \mathcal{D}_k}, Y_i > w_n] &= \widehat{u}_{j,k}^\top \widehat{u}_{j,k} \left(1 + O\left(w_n^{-\beta}\right)\right)\end{aligned}$$

from Lemma 3. Then, we can apply Lindeberg's CLT. In particular, the Lindeberg condition is satisfied by checking, for any  $\delta > 0$ ,

$$\lim_{n_0 \rightarrow \infty} \frac{1}{n_0} \sum_{k=1}^K \sum_{i \in \mathcal{D}_k} \mathbb{E} [\Psi_{i,k}^2 \mathbf{1}\{\|\Psi_{i,k}\|/\sqrt{n_k} \geq \delta\} | Y_i > w_n] = 0.$$

This is obtained by Lemma 3 (v), since

$$\begin{aligned}\mathbb{E} [\Psi_{i,k}^4 | Y_i > w_n] &= 25 \mathbb{E} \left[ v_j^{-2} \left| \widehat{u}_{j,k}^\top X_i \right|^4 \left(1 + O\left(w_n^{-\beta(X_i)}\right)\right) \middle| Y_i > w_n \right] \\ &\leq C (\gamma_{2n_0})^4 = O(\log^2 n_0) = o(\sqrt{n_k}),\end{aligned}$$

where the inequality follows from that  $\max_{1 \leq i \leq n_k} |\widehat{u}_{j,k}^\top X_i| \leq \gamma_{2n_0}$ , and the last equality follows from the fact that  $n_0 = Kn_k$  with a fixed  $K$ . Thus Lindeberg's CLT applies.  $\square$

#### A.4 Proof of Theorem 4

*Proof.* Substitute the definition of  $\tilde{\theta}_j^{\text{on}}$  yields that

$$\begin{aligned}\tilde{\theta}_j^{\text{on}} - \theta_{0j} &= \widehat{\theta}_j^{\text{on}} - \theta_{0j} + \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} \left\{ \exp\left(X_t^\top \widehat{\theta}^{\text{on}}\right) \log(Y_t/\bar{w}) - 1 \right\} X_t \\ &= \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} \left\{ \exp(X_t^\top \theta_0) \log(Y_t/\bar{w}) - 1 \right\} X_t \\ &\quad + \left( \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} \log(Y_t/\bar{w}) \left\{ \exp(X_t^\top \theta_0) \right\} X_t X_t^\top - e_j \right) (\widehat{\theta}^{\text{on}} - \theta_0) \\ &\quad + \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} \log(Y_t/\bar{w}) X_t \left\{ \exp\left(X_t^\top \widehat{\theta}^{\text{on}} + t X_t^\top (\theta_0 - \widehat{\theta}^{\text{on}})\right) \right\} \left[ X_t^\top (\widehat{\theta}^{\text{on}} - \theta_0) \right]^2 \\ &\equiv A_1 + A_2 + A_3,\end{aligned}$$

for some  $t \in (0, 1)$ .

Note that  $\widehat{\Xi}$  is a function of  $X$  only, and hence, conditional on  $X$ ,  $Z_t(\theta_0)$  is i.i.d. Also, conditional on  $Y_t > \bar{w}$  and  $X_t = x$ ,  $\log(Y_t/\bar{w})$  is exponentially distributed with parameter  $\exp(x^\top \theta_0)$ , yielding that

$$\begin{aligned}\mathbb{E} [\{\exp(X_t^\top \theta_0) \log(Y_t/\bar{w}) - 1\} X_t | Y_t > \bar{w}] &= 0 \\ \mathbb{E} \left[ \{\exp(X_t^\top \theta_0) \log(Y_t/\bar{w}) - 1\}^2 X_t X_t^\top | Y_t > \bar{w} \right] &= \Sigma_{\bar{w}}.\end{aligned}$$

Finally,  $(\widehat{\Xi} \widehat{\Sigma} \widehat{\Xi}^\top)_{j,j} \xrightarrow{p} (\Xi \Sigma_{\bar{w}} \Xi^\top)_{j,j}$  follows from the proof of Chen et al. (2020, Theorem 5.2). In particular, we need to check their Lemmas E.1 and E.2. Note that our  $X_t$  has bounded support in all component, implying that  $X_t$  is a sub-Gaussian vector. Then, for all  $j = 1, \dots, p$ ,

$$\sqrt{T_0} A_1 \xrightarrow{d} \mathcal{N}(0, (\Xi \Sigma_{\bar{w}} \Xi^\top)_{j,j}).$$

It remains to show that  $A_2 = o_p(\sqrt{T_0})$  and  $A_3 = o_p(\sqrt{T_0})$ .

We study  $A_2$ . Assumption 4 yields that  $\{(\exp(X_t^\top \theta_0) \log(Y_t/\bar{w}) - 1) X_t X_t^\top\}$  are subexponential random variables conditional on  $X_t$  and  $Y_t > \bar{w}$ . Then, applying the concentration inequality for subexponential random variables (e.g., Vershynin, 2010, Proposition 5.16), we have that, with probability at least  $1 - p^{-c}$ ,

$$\left\| \frac{1}{T_0} \sum_{t=1}^{T_0} (\log(Y_t/\bar{w}) \exp(X_t^\top \theta_0) - 1) X_t X_t^\top \right\|_\infty \lesssim \sqrt{(\log p)/n_0}.$$

Furthermore, Chen et al. (2020, Lemma E.2) establishes that  $\widehat{\Xi}_j \xrightarrow{p} \Xi_{\bar{w}_j}$ , which satisfies that  $\|\Xi_{\bar{w}_j}\|_\infty < \infty$  from the assumption  $C_2^{-1} \leq \lambda_{\min}(\Sigma_{\bar{w}}) \leq \lambda_{\max}(\Sigma_{\bar{w}}) \leq C_2$ . Then, we have

$$\begin{aligned}\sqrt{T_0} |A_2| &\leq \left\| \frac{\widehat{\Xi}_j^\top}{T_0} \sum_{t=1}^{T_0} (\log(Y_t/\bar{w}) \exp(X_t^\top \theta_0) - e_j) X_t X_t^\top \right\| \cdot \left\| \widehat{\theta}^{\text{on}} - \theta_0 \right\|_1 \\ &= O_p \left( \sqrt{(\log p)/n_0} \cdot \left\| \widehat{\theta}^{\text{on}} - \theta_0 \right\|_1 \right) \\ &\stackrel{(1)}{=} O_p \left( \frac{s_0 \log p}{n_0} \right) \stackrel{(2)}{=} o_p(1),\end{aligned}$$

where (1) follows from Lemma 7, and (2) from the condition of the theorem.

For  $A_3$ , define

$$\Delta_t = \log(Y_t/\bar{w}) \left\{ \exp \left( X_t^\top \widehat{\theta}^{\text{on}} + t X_t^\top (\theta_0 - \widehat{\theta}^{\text{on}}) \right) \right\} \cdot \left[ X_t^\top (\widehat{\theta}^{\text{on}} - \theta_0) \right]^2.$$

A similar argument as in the proof of Lemma 5 establishes that

$$\frac{1}{T_0} \sum_{t=1}^{T_0} \left( \Delta_t - \left[ X_t^\top (\widehat{\theta}^{\text{on}} - \theta_0) \right]^2 \right) = o_p(1).$$

It follows that

$$\begin{aligned}
\sqrt{T_0} |A_3| &= \left| \frac{1}{\sqrt{T_0}} \sum_{t=1}^{T_0} \widehat{\Xi}_j^\top X_t \Delta_t \right| \\
&\leq \max_{1 \leq t \leq T_0} |\widehat{\Xi}_j^\top X_t| \cdot \left| \frac{1}{\sqrt{T_0}} \sum_{t=1}^{T_0} \Delta_t \right| \\
&\stackrel{(1)}{\lesssim} \frac{\sqrt{\log T_0}}{\sqrt{T_0}} \sum_{t=1}^{T_0} \left[ X_t^\top (\widehat{\theta}^{\text{on}} - \theta_0) \right]^2 \\
&= \frac{\sqrt{\log T_0}}{\sqrt{T_0}} (\widehat{\theta}^{\text{on}} - \theta_0)^\top X^\top X (\widehat{\theta}^{\text{on}} - \theta_0) \\
&\lesssim \sqrt{T_0} \gamma_{2n_0} \left\| \widehat{\theta}^{\text{on}} - \theta_0 \right\|_2^2 \\
&\stackrel{(2)}{\lesssim} \frac{\sqrt{\log T_0} s_0 \log p}{\sqrt{T_0}} \\
&\stackrel{(3)}{=} o_p(1),
\end{aligned}$$

where (1) is by  $\max_{1 \leq t \leq T_0} |\widehat{\Xi}_j^\top X_t| \lesssim \sqrt{\log T_0}$ , which further follows from that  $X_t$  has compact support and hence sub-Gaussian; (2) is by Lemma 6; and (3) follows from the condition in the theorem.  $\square$

## B Useful Lemmas

**Lemma 1.** Define

$$Z_{n_0,i,j} = [\alpha(X_i) \log(Y_i/w_n) - 1] X_{i,j}.$$

Suppose Assumptions 1-3 hold. Then for all  $u$ ,

$$\mathbb{P}(|Z_{n_0,i,j}| > u | Y_i > w_n) \leq C_1 \exp(-C_2 u)$$

for some finite constants  $C_1, C_2$ , which do not depend on  $w_n$ .

*Proof of Lemma 1.* On the event that  $\{X_{i,j} = 0\}$ ,  $Z_{n_0,i,j} = 0$  and hence the lemma follows trivially.

Now consider the event  $\{X_{i,j} \neq 0\}$ . For any  $u > 0$ ,

$$\begin{aligned}
& \mathbb{P}(|Z_{n_0,i,j}| > u | Y_i > w_n) \\
&= \mathbb{E}[\mathbb{P}(|Z_{n_0,i,j}| > u | X_i, Y_i > w_n) | Y_i > w_n] \\
&= \mathbb{E}\left[\mathbb{P}\left(\alpha(X_i) \log(Y_i/w_n) > 1 + \frac{u}{|X_{i,j}|} \middle| X_i, Y_i > w_n\right)\right] \\
&\quad + \mathbb{E}\left[\mathbb{P}\left(\alpha(X_i) \log(Y_i/w_n) < 1 - \frac{u}{|X_{i,j}|} \middle| X_i, Y_i > w_n\right)\right] \\
&= P_1(u) + P_2(u).
\end{aligned}$$

For  $P_1(u)$ , Assumption 1 implies that for any  $x \in \mathbb{R}^{\dim\{X\}}$ ,

$$\begin{aligned}
& \mathbb{P}\left(\alpha(x) \log(Y/w_n) > \left(1 + \frac{u}{|x_j|}\right) \middle| X = x, Y > w_n\right) \\
&= \mathbb{P}\left(\frac{Y}{w_n} > \exp\left(\frac{1}{\alpha(x)}\left(1 + \frac{u}{|x_j|}\right)\right) \middle| X = x, Y > w_n\right) \\
&= e^{-(1+u/|x_j|)} \left(1 + \frac{c_1(x)}{c_0(x)} w_n^{-\beta(x)} \left(\exp\left(-\frac{\beta(x)}{\alpha(x)}\left(1 + \frac{u}{|x_j|}\right)\right) - 1\right) + o(w_n^{-\beta(x)})\right) \\
&\leq e^{-(1+u/|x_j|)} \left(1 + O(w_n^{-\beta(x)})\right)
\end{aligned}$$

where  $x_j$  denote the  $j$ th component of the vector  $x$ . Given that  $|X_{i,j}|$  has a bounded support, we proceed with  $|X_{i,j}| \leq 1$  without loss of generality. Let  $C$  denote a generic constant, whose value could change line-by-line. It follows that

$$\begin{aligned}
P_1(u) &= \int_0^1 e^{-(1+u/x)} \left(1 + O(w_n^{-\beta(x)})\right) f_{X_{i,j}|Y_i>w_n}(x) dx \\
&\leq \bar{f} \int_0^1 e^{-(1+u/x)} dx \left(1 + O(w_n^{-\beta(x)})\right) \\
&\leq C_1 e^{-C_2 u},
\end{aligned}$$

where the first inequality is from  $f_{X_{i,j}|Y_i>w_n} < \bar{f}$  (Assumption 2), and the second inequality is by direct calculation with  $C_1 = 2\bar{f}e^{-1}$  and  $C_2 = 1$ .

For  $P_2(u)$ , since  $\alpha(X_i) > 0$  and  $Y_i > w_n$ ,

$$P_2(u) \leq \mathbb{P}(|X_{i,j}| > u | Y_i > w_n).$$

The fact that  $|X_{i,j}| \leq 1$  (conditional on  $Y_i$ ) implies that  $X_{i,j}$  is sub-Gaussian and also sub-exponential, which further implies that  $P_2(u) \leq C_1 e^{-C_2 u}$  with  $C_1 = 2$  and  $C_2 = \log 2$ . The proof is complete by combining  $P_1(u)$  and  $P_2(u)$  and setting  $\bar{u} = 1$ .  $\square$

**Lemma 2.** Suppose that the conditions of Theorem 2 hold. With probability at least  $1 - p^{-c} - n_0^{-c}$ , there exists  $\hat{u}_{j,k}$  such that for each  $j = 1, \dots, p$  and  $k = 1, \dots, K$ ,

$$\begin{aligned} \left\| \frac{1}{K} \sum_{k=1}^K \left( \frac{\hat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} \exp(X_i^\top \theta_0) \log(Y_i/w_n) X_i X_i^\top - e_j \right) \right\|_\infty &\leq \gamma_{1n_0} \\ \max_{i \in \mathcal{D}_k^c} |X_i^\top \hat{u}_{j,k}| &\leq \gamma_{2n_0}. \end{aligned} \quad (\text{B.1})$$

*Proof of Lemma 2.* Note that  $\hat{u}_{j,k}$  is constructed to satisfy

$$\begin{aligned} \left\| \frac{1}{K} \sum_{k=1}^K \left( \frac{\hat{u}_{j,k}^\top}{n_k} \sum_{i=1}^{n_k} X_i X_i^\top - e_j \right) \right\|_\infty &\leq \gamma_{1n_0} \\ \max_{i \in \mathcal{D}_k} |X_i^\top \hat{u}_{j,k}| &\leq \gamma_{2n_0}. \end{aligned} \quad (\text{B.2})$$

Therefore, we show that (i) such  $\hat{u}_{j,k}$  exists with probability at least  $1 - p^{-c} - n_0^{-c}$ , and (ii) such  $\hat{u}_{j,k}$  also satisfies the constraint (B.1) in the statement. For (i), we establish the following two steps:

- (a) the matrix  $\Sigma_{w_n} := \mathbb{E}[X_i X_i^\top | Y_i > w_n]$  is invertible;
- (b) the  $j$ -th column  $\tilde{\mu}_{j,k}$  of  $\Sigma_{w_n}^{-1}$ ,  $j = 1, \dots, p$ , is feasible for the constraint in (B.2) with probability at least  $1 - p^{-c} - n_0^{-c}$ .

Step (a) is guaranteed by our parameter space assumption that  $M^{-1} \leq \lambda_{\min}(\Sigma_{w_n}) \leq \lambda_{\max}(\Sigma_{w_n}) \leq M$  for all  $n$ . Step (b) follows from the argument of proving (1.2) in Cai et al. (2023) with their  $h(\cdot) = 1$  and conditioning on  $\{Y_i > w_n\}$ .

We now focus on establishing (ii), which is equivalent to showing

$$\left\| \frac{1}{K} \sum_{k=1}^K \hat{u}_{j,k}^\top \left( \frac{1}{n_k} \sum_{i=1}^{n_k} (\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1) X_i X_i^\top \right) \right\|_\infty \leq \gamma_{1n_0}. \quad (\text{B.3})$$

To this end, note that  $\{(\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1) X_i X_i^\top\}_{i \in \mathcal{D}_k}$  are i.i.d. and satisfy

$$\begin{aligned} &\mathbb{E}[(\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1) X_i X_i^\top | Y_i > w_n] \\ &= \mathbb{E}\left[X_i X_i^\top \left(1 + O\left(w_n^{-\beta(X_i)}\right)\right) | Y_i > w_n\right] \\ &\leq \mathbb{E}[X_i X_i^\top | Y_i > w_n] \left(1 + O\left(w_n^{-\beta}\right)\right). \end{aligned}$$

Moreover, Assumption 1 and Lemma 1 yield that  $\{(\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1) X_i X_i^\top\}_{i \in \mathcal{D}_k}$  are subexponential random variables conditional on  $X_i$  and  $Y_i > w_n$ . Then applying the concentration inequality

for subexponential random variables (e.g., Vershynin, 2010, Proposition 5.16) and the condition that  $w_n^{-\beta} \lesssim \gamma_{1n_0}$  (Assumption 3), we have that with probability at least  $1 - p^{-c}$ .

$$\left\| \frac{1}{n_k} \sum_{i=1}^{n_k} (\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1) X_i X_i^\top \right\|_\infty \lesssim \gamma_{1n_0}$$

for  $\gamma_{1n_0} \asymp \sqrt{(\log p)/n_0}$ . From Step (b) above, we know that the  $j$ -th column  $\tilde{\mu}_{j,k}$  of  $\Sigma_{w_n}^{-1}$  is a valid solution for  $\hat{\mu}_{j,k}$  and satisfies that  $\|\tilde{\mu}_{j,k}\|_\infty < \infty$  from the assumption  $C_2^{-1} \leq \lambda_{\min}(\Sigma_{w_n}) \leq \lambda_{\max}(\Sigma_{w_n}) \leq C_2$  for all  $n$ . Then (B.3) follows since  $K$  is finite. Thus, the conclusion follows from (i) and (ii).  $\square$

**Lemma 3.** *Suppose Assumptions 1-3 hold. Then the following holds.*

- (i)  $\mathbb{E}[\log(Y_i/w_n) | X_i, Y_i > w_n] = \frac{1}{\exp(X_i^\top \theta_0)} \left( 1 + O(w_n^{-\beta(X_i)}) \right)$
- (ii)  $\|\mathbb{E}[\{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i | Y_i > w_n]\|_\infty \leq C w_n^{-\beta}$
- (iii)  $\left\| \mathbb{E} \left[ X_i X_i^\top \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\}^2 | Y_i > w_n \right] - \mathbb{E}[X_i X_i^\top | Y_i > w_n] \right\|_\infty \leq C w_n^{-\beta}$
- (iv)  $\|\mathbb{E}[X_i X_i^\top \exp(X_i^\top \theta_0) \log(Y_i/w_n) | Y_i > w_n] - \Sigma_{w_n}\|_\infty \leq C w_n^{-\beta}$
- (v)  $\mathbb{E} \left[ \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\}^4 | X_i, Y_i > w_n \right] = 25 \left( 1 + O(w_n^{-\beta(X_i)}) \right).$

*Proof of Lemma 3.* Assumption 1 implies that conditional  $X_i = x$ ,

$$\begin{aligned} & \mathbb{P}(Y_i > w_n | X_i = x) \\ &= w_n^{-\alpha(x)} \mathcal{L}(w_n; x) \\ &= w_n^{-\alpha(x)} \left[ c_0(x) + c_1(x) w_n^{-\beta(x)} + r(w_n, x) \right]. \end{aligned}$$

Accordingly, the PDF of  $Y_i$  conditional on  $X_i = x$  and  $Y_i > w_n$  satisfies

$$\begin{aligned} & f_{Y|Y>w_n, X=x}(y) \\ &= \frac{f_{Y|X=x}(y)}{\mathbb{P}(Y_i > w_n | X_i = x)} \\ &= \frac{c_0(x) \alpha(x) y^{-\alpha(x)-1} + c_1(x) \beta(x) y^{-\alpha(x)-\beta(x)-1}}{w_n^{-\alpha(x)} \left[ c_0(x) + c_1(x) w_n^{-\beta(x)} + r(w_n, x) \right]} (1 + o(1)) \\ &= \alpha(x) (y/w_n)^{-\alpha(x)} y^{-1} \left( \frac{c_0(x) + \alpha(x)^{-1} c_1(x) \beta(x) y^{-\beta(x)}}{c_0(x) + c_1(x) w_n^{-\beta(x)} + r(w_n, x)} \right) (1 + o(1)) \\ &= \alpha(x) (y/w_n)^{-\alpha(x)} y^{-1} (1 + \Delta(w_n; x)), \end{aligned} \tag{B.4}$$

where the higher-order term  $\Delta(w_n; x)$  satisfies that

$$|\Delta(w_n; x)| \leq \left| \frac{c_1(x)}{c_0(x)} \frac{\beta(x)}{\alpha(x)} \right| w_n^{-\beta(x)}. \quad (\text{B.5})$$

Using (B.4), we have that

$$\begin{aligned} & \mathbb{E} [\log(Y_i/w_n) | X_i = x, Y_i > w_n] \\ &= \int_{w_n}^{\infty} \log\left(\frac{y}{w_n}\right) f_{Y|Y>w_n, X=x}(y) dy \\ &= \int_1^{\infty} \log(t) t^{-\alpha(x)-1} dt \times \alpha(x) \left(1 + O\left(w_n^{-\beta(x)}\right)\right) \\ &= \frac{1}{\alpha(x)} (1 + \Delta(w_n; x)), \end{aligned}$$

where the second equality is by the change of variable  $y/w_n \rightarrow t$ . Part (i) follows from that  $\alpha(x) = \exp(x^\top \theta_0)$  and (B.5).

For Part (ii), (B.4) yields that

$$\begin{aligned} & \|\mathbb{E}[\{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i | Y_i > w_n]\|_\infty \\ &= \|\mathbb{E}[X_i \Delta(w_n; X_i)]\|_\infty \\ &\leq \left\| \mathbb{E} \left[ X_i \times \left| \frac{c_1(X_i)}{c_0(X_i)} \frac{\beta(X_i)}{\alpha(X_i)} \right| \times w_n^{-\beta(X_i)} | Y_i > w_n \right] \right\|_\infty \\ &\leq C w_n^{-\frac{\beta}{2}} \|\mathbb{E}[X_i | Y_i > w_n]\|_\infty. \end{aligned}$$

For Part (iii), (B.4) yields that

$$\begin{aligned} & \mathbb{E} [\log(Y_i/w_n)^2 | X_i = x, Y_i > w_n] \\ &= \int_{w_n}^{\infty} \left( \log\left(\frac{y}{w_n}\right) \right)^2 f_{Y|Y>w_n, X=x}(y) dy \\ &= \int_1^{\infty} \log(t)^2 t^{-\alpha(x)-1} dt \times \alpha(x) \left(1 + O\left(w_n^{-\beta(x)}\right)\right) \\ &= \frac{2}{\alpha(x)^2} (1 + \Delta(w_n; x)). \end{aligned}$$

Then

$$\begin{aligned}
& \left\| \mathbb{E} \left[ X_i X_i^\top \{ \exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1 \}^2 | Y_i > w_n \right] - \mathbb{E} [X_i X_i^\top | Y_i > w_n] \right\|_\infty \\
&= \| \mathbb{E} \left[ X_i X_i^\top \exp(2X_i^\top \theta_0) \log(Y_i/w_n)^2 | Y_i > w_n \right] \\
&\quad - 2\mathbb{E} [X_i X_i^\top \exp(X_i^\top \theta_0) \log(Y_i/w_n) | Y_i > w_n] \|_\infty \\
&\leq 4\mathbb{E} [X_i X_i^\top |\Delta(w_n; x)| | Y_i > w_n] \\
&\leq Cw_n^{-\beta} \|\mathbb{E} [X_i X_i^\top | Y_i > w_n]\|_\infty.
\end{aligned}$$

For Part (iv),

$$\begin{aligned}
& \| \mathbb{E} [X_i X_i^\top \exp(X_i^\top \theta_0) \log(Y_i/w_n) | Y_i > w_n] - \Sigma_{w_n} \|_\infty \\
&= \| \mathbb{E} [X_i X_i^\top \{ \exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1 \} | Y_i > w_n] \|_\infty \\
&\leq Cw_n^{-\beta} \|\mathbb{E} [X_i X_i^\top | Y_i > w_n]\|_\infty,
\end{aligned}$$

which follows from Part (ii).

For Part (v),

$$\begin{aligned}
& \mathbb{E} \left[ \{ \exp(x^\top \theta_0) \log(Y_i/w_n) - 1 \}^4 | X_i = x, Y_i > w_n \right] \\
&= \mathbb{E} \left[ \alpha(x)^4 \log(Y_i/w_n)^4 | X_i = x, Y_i > w_n \right] \\
&\quad - 4\mathbb{E} \left[ \alpha(x)^3 \log(Y_i/w_n)^3 | X_i = x, Y_i > w_n \right] \\
&\quad + 6\mathbb{E} \left[ \alpha(x)^2 \log(Y_i/w_n)^2 | X_i = x, Y_i > w_n \right] \\
&\quad - 4\mathbb{E} [\alpha(x) \log(Y_i/w_n) | X_i = x, Y_i > w_n] + 1.
\end{aligned}$$

For the first item above, the same argument as part (i) yields that

$$\begin{aligned}
& \mathbb{E} \left[ \alpha(x)^4 \log(Y_i/w_n)^4 | X_i = x, Y_i > w_n \right] \\
&= \int_{w_n} \alpha(x)^4 \log \left( \frac{y}{w_n} \right)^4 f_{Y|Y>w_n, X=x}(y) dy \\
&= \int_1 \alpha(x)^5 \log(t)^4 t^{-\alpha(x)-1} dt \left( 1 + O(w_n^{-\beta(x)}) \right) \\
&= 24 \left( 1 + O(w_n^{-\beta(x)}) \right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{E} \left[ \alpha(x)^3 \log(Y_i/w_n)^3 | X_i = x, Y_i > w_n \right] &= 6 \left( 1 + O(w_n^{-\beta(x)}) \right). \\ \mathbb{E} \left[ \alpha(x)^2 \log(Y_i/w_n)^2 | X_i = x, Y_i > w_n \right] &= 2 \left( 1 + O(w_n^{-\beta(x)}) \right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E} \left[ \{\exp(x^\top \theta_0) \log(Y_i/w_n) - 1\}^4 | X_i = x, Y_i > w_n \right] \\ = 25 \left( 1 + O(w_n^{-\beta(x)}) \right),\end{aligned}$$

which completes the proof.  $\square$

**Lemma 4.** *Under the assumptions of Theorem 1, it holds that for some constants  $c$  and  $T$ ,*

$$\mathbb{P} \left( \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2=1} \frac{1}{n_0} \sum_{i=1}^{n_0} \varphi_T [\log(Y_i/w_i)] \varphi_T [(b^\top X_i)^2] > c \middle| \{Y_i > w_n\}_{i=1}^{n_0} \right) \geq 1 - e^{-cn_0 - \log p},$$

*Proof of Lemma 4.* The proof follows similarly as the proof of Cai et al. (2023, Lemma 4). Define

$$g_b(y, x) = \varphi_T [\log(y)] \varphi_T [(b^\top x)^2].$$

We need to show

- (i)  $\mathbb{E} [g_b(Y_i/w_n, (b^\top X_i)^2) | Y_i > w_n] \geq c/2$  for some universal constant  $c > 0$ , and
- (ii) For the random variable

$$Z(t) = \inf_{b \in \mathcal{C}(\varsigma, S), \|b\|_2=1, \|b\|_1=t} \left| \frac{\frac{1}{n_0} \sum_{i=1}^{n_0} g_b(Y_i/w_n, (b^\top X_i)^2)}{\mathbb{E} [g_b(Y_i/w_n, (b^\top X_i)^2) | Y_i > w_n]} - 1 \right|,$$

it holds that

$$\mathbb{P} \left( Z(t) \geq c/4 + C \sqrt{\frac{\log p}{n_0}} t \middle| \{Y_i > w_n\}_{i=1}^{n_0} \right) \leq c_1 \exp(-c_2 - c_3 t^2 \log p).$$

To show (i), note that on the set  $\|b\|_2 = 1$ , Lemma 1 implies that

$$\begin{aligned}\mathbb{E} [\log(Y_i/w_n) (b^\top X_i)^2 | Y_i > w_n] \\ = \mathbb{E} [(b^\top X_i)^2 \mathbb{E} [\log(Y_i/w_n) | Y_i > w_n, X_i = x] | Y_i > w_n] \\ = \mathbb{E} \left[ \frac{(b^\top X_i)^2}{\alpha(X_i)} (1 + \Delta(w_n; X_i)) \middle| Y_i > w_n \right] \\ \geq \mathbb{E} [(b^\top X_i)^2] / \underline{\alpha} > c > 0.\end{aligned}$$

Then it suffices to show that

$$\begin{aligned}
& c/2 \\
& \geq \mathbb{E} \left[ \log(Y_i/w_n) (b^\top X_i)^2 | Y_i > w_n \right] - \mathbb{E} \left[ \varphi_T [\log(Y_i/w_n)] \varphi_T [(b^\top X_i)^2] | Y_i > w_n \right] \\
& = \mathbb{E} \left[ (\log(Y_i/w_n) - \varphi_T [\log(Y_i/w_n)]) (b^\top X_i)^2 | Y_i > w_n \right] \\
& \quad + \mathbb{E} \left[ \varphi_T [\log(Y_i/w_n)] \left( (b^\top X_i)^2 - \varphi_T [(b^\top X_i)^2] \right) | Y_i > w_n \right] \\
& \equiv A_1 + A_2.
\end{aligned}$$

For  $A_1$ , the proof of Lemma 3.(i) implies that

$$\begin{aligned}
& \mathbb{E} [\log(Y_i/w_n) \cdot 1 \{\log(Y_i/w_n) > T\} | Y_i > w_n, X_i = x] \\
& = \int_{w_n \exp(T)}^{\infty} \log \left( \frac{y}{w_n} \right) f_{Y|Y>w_n, X=x}(y) dy \\
& = \int_{\exp(T)}^{\infty} \log(t) t^{-\alpha(x)-1} dt \times \alpha(x) \left( 1 + O(w_n^{-\beta(x)}) \right) \\
& = \exp(-\alpha(x)T) \frac{(2 + \alpha(x)T)(2 + \alpha(x)T)}{\alpha(x)^3} \left( 1 + O(w_n^{-\beta(x)}) \right) \\
& \leq c_0 T^2 \exp(-c_1 T).
\end{aligned}$$

Therefore,

$$A_1 \leq c_0 T^2 \exp(-c_1 T) \mathbb{E} [(b^\top X_i)^2 | Y_i > w_n] \leq c_0 T^2 \exp(-c_1 T),$$

which is bounded by  $c/2$  by setting a sufficiently large  $T$ .

For  $A_2$ , since  $X_{i,j}$  has a bounded support for all  $j$  (Assumption 2), it implies that  $X_i$  is sub-Gaussian vector. Then, we can use Cauchy-Schwartz inequality and the fact that  $\varphi_T(x) \leq T$  to obtain that

$$\begin{aligned}
A_2 & \leq T \mathbb{E} \left[ \left( (b^\top X_i)^2 - \varphi_T [(b^\top X_i)^2] \right) | Y_i > w_n \right] \\
& \leq T \mathbb{E} \left[ (b^\top X_i)^2 \cdot 1 \left[ (b^\top X_i)^2 > T \right] | Y_i > w_n \right] \\
& \leq T \sqrt{\mathbb{E} \left[ (b^\top X_i)^4 | Y_i > w_n \right]} \mathbb{P}^{1/2} \left( (b^\top X_i)^2 > T \right) \\
& \leq c_0 T^2 \exp(-c_2 T),
\end{aligned}$$

which is again bounded by  $c/2$  by setting a sufficiently large  $T$ . Then (i) is established by combining  $A_1$  and  $A_2$ .

For (ii), the truncation function  $\varphi_T(\cdot)$  yields that  $\|g_b(y, x)\|_\infty \leq T^2$ . The rest of the proof follows similarly from the proof of (2.11) in Cai et al. (2023).  $\square$

**Lemma 5.** Suppose the conditions of Theorem 3 hold. Then it holds that

$$\frac{1}{n_0} \sum_{k=1}^K \sum_{i=1}^{n_k} \left( \Delta_i - \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \right) = o_p(1).$$

*Proof of Lemma 5.* Recall the definition

$$\begin{aligned} \Delta_i &= \log(Y_i/w_n) \exp \left( X_i^\top \widehat{\theta}_k + t X_i^\top (\theta_0 - \widehat{\theta}_k) \right) \cdot \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \\ &= \log(Y_i/w_n) \exp(X_i^\top \theta_0) \Xi_i \cdot \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \end{aligned}$$

for some  $t \in (0, 1)$ , where

$$\Xi_i \equiv \exp \left( X_i^\top (\widehat{\theta}_k - \theta_0) (1-t) \right).$$

Since  $X_i$  is a sub-Gaussian vector, Theorem 1 implies that

$$\begin{aligned} \Xi_i &\leq \exp \left( C \left\| \widehat{\theta}_k - \theta_0 \right\|_2^2 \right) \\ &\leq \exp \left( C \frac{s_0 \log p}{n_0} \right) \\ &\leq (1 + C \frac{s_0 \log p}{n_0}), \end{aligned}$$

where the last inequality follows from the fact that  $e^x \leq 1 + 3x$  for  $x \in (0, 1)$  and  $\frac{s_0 \log p}{n_0} \rightarrow 0$  (Assumption 3).

Since  $\widehat{\theta}_k$  is constructed using the subsample  $\mathcal{D}_k^c$ , Lemma 3 yields that for  $i \in \mathcal{D}_k$ ,  $\{\Delta_i\}$  are i.i.d. and satisfy that

$$\begin{aligned} &\mathbb{E} \left[ \left( \Delta_i - \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \right)^2 \middle| \mathcal{D}_k^c, X_i, Y_i > w_n \right] \\ &= \mathbb{E} \left[ (\log(Y_i/w_n) \exp(X_i^\top \theta_0) \Xi_i - 1)^2 \middle| X_i, Y_i > w_n \right] \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^4 \\ &\leq (\Xi_i^2 - 2\Xi_i + 1) \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^4 \left( 1 + O(w_n^{-\beta(X_i)}) \right) \\ &\leq C \left( \frac{s_0 \log p}{n_0} \right) \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^4 \left( 1 + O(w_n^{-\beta}) \right). \end{aligned}$$

Then by Cauchy-Schwartz inequality and integrating over  $X_i$ , we have that

$$\begin{aligned} & \mathbb{E} \left[ \left| \Delta_i - \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \right| \middle| \mathcal{D}_k^c, X_i, Y_i > w_n \right] \\ & \leq C \left( \frac{s_0 \log p}{n_0} \right)^{1/2} \mathbb{E} \left[ \left[ X_i^\top (\widehat{\theta}_k - \theta_0) \right]^2 \middle| \mathcal{D}_k^c, Y_i > w_n \right] \left( 1 + O(w_n^{-\beta}) \right) \\ & = o(1). \end{aligned}$$

Then the result follows from Markov's inequality.  $\square$

**Lemma 6.** *Under the assumptions of Theorem 4, it holds that*

$$\left\| \widehat{\theta}^{on} - \theta_0 \right\|_2 \lesssim \sqrt{\frac{s_0(\log p)}{T_0}} \text{ and } \left\| \widehat{\theta}^{on} - \theta_0 \right\|_1 \lesssim \sqrt{\frac{s_0^2(\log p)}{T_0}}.$$

*Proof of Lemma 6.* This result follows from Proposition 1 and Lemma 1 in Agarwal et al. (2012). To apply these results, we introduce some notations and describe their relations with those in Agarwal et al. (2012). Denote  $C$  as a generic universal constant, whose value may vary across lines.

First, the loss function  $\bar{\mathcal{L}}(\theta)$  in Agarwal et al. (2012) becomes

$$\begin{aligned} \bar{\mathcal{L}}(\theta) &= \mathbb{E} [\{\exp(X_t^\top \theta) + 1\} \log(Y_t/\bar{w}) - X_t^\top \theta | Y_t > \bar{w}] \\ &= \mathbb{E} [\{\exp(X_t^\top \theta) + 1\} \exp(X_t^\top \theta_0) - X_t^\top \theta | Y_t > \bar{w}], \end{aligned}$$

which is satisfies their Assumptions 1 (locally Lipschitz) and 2' (locally restricted strong convexity).

Second, denote the stochastic gradient as

$$g_t(\theta) = \{\exp(X_t^\top \theta) \log(Y_t/\bar{w}) - 1\} X_t,$$

where recall that we use only the tail data  $Y_t > \bar{w}$ . Note that  $g_t(\theta)$  is only sub-exponential (Lemma 1) instead of sub-Gaussian, which is the key difference from Agarwal et al. (2012). Define

$$e_t(\theta) = g_t(\theta) - \mathbb{E}[g_t(\theta)].$$

Instead of bounding  $\mathbb{E}[\exp(\|e_t(\theta)\|_\infty)]$ , we now bound  $\mathbb{E}[\|e_t(\theta)\|_\infty^4]$ . Since all components of  $X_t$  have a compact support, it holds that for some constant  $C$

$$\|e_t(\theta)\|_\infty^4 \leq C \{\exp(X_t^\top \theta) \log(Y_t/\bar{w}) - 1\}^4.$$

Then, some calculation yields that for any  $\theta$  such that  $\|\theta - \theta_0\|_1 \leq R$ ,

$$\begin{aligned}
& \mathbb{E} \left[ \|e_t(\theta)\|_\infty^4 \right] \\
& \leq C \mathbb{E} \left[ \{\exp(X_t^\top \theta) \log(Y_t/\bar{w}) - 1\}^4 \right] \\
& = C \{ \mathbb{E} [\exp(4X_t^\top(\theta - \theta_0))] - 4\mathbb{E} [\exp(3X_t^\top(\theta - \theta_0))] \\
& \quad + 6\mathbb{E} [\exp(2X_t^\top(\theta - \theta_0))] - 4\mathbb{E} [\exp(X_t^\top(\theta - \theta_0))] \} + 1 \\
& \leq C \{ \exp\{16\|\theta - \theta_0\|_2^2 C^2/2\} + 4 \exp\{9\|\theta - \theta_0\|_2^2 C^2/2\} \\
& \quad + 6 \exp\{4\|\theta - \theta_0\|_2^2 C^2/2\} + 4 \exp\{\|\theta - \theta_0\|_2^2 C^2/2\} \} \\
& \leq 16C \exp\{8R^2C^2\} + 1,
\end{aligned} \tag{B.6}$$

$$16C \exp\{8R^2C^2\} + 1, \tag{B.7}$$

where (B.6) is from the fact that  $X$  has bounded support implies that it is sub-Gaussian. Accordingly, set

$$\sigma^2(R) = \sqrt{(16C \exp\{8R^2C^2\} + 1)},$$

yielding that  $\mathbb{E} \left[ \|e_t(\theta)\|_\infty^4 \right] \leq \sigma^4(R)$ .

Third, by carefully examining the proof of Proposition 1 in Agarwal et al. (2012), sub-Gaussianity is only required in their Lemma 7. Therefore, we establish Lemma 7 below, which is a weaker version of their Lemma 7. Then using Proposition 1 and Lemma 1 in Agarwal et al. (2012), we have that

$$\left\| \widehat{\theta}_i - \theta_0 \right\|_2 \lesssim \sqrt{s_0} \lambda_i \text{ and } \left\| \widehat{\theta}_i - \theta_0 \right\|_1 \lesssim s_0 \lambda_i, \tag{B.8}$$

where  $\widehat{\theta}_i$  denotes the estimates in the  $i$ -th epoch. Note that their  $\|\theta_{S^c}^*\|_1 = 0$  given our sparsity condition.

By setting the regularization parameter  $\lambda_i$  as in eq.(34) in Agarwal et al. (2012), we have that

$$\lambda_i^2 = \frac{R_i C_1^{-1}}{s_0 \sqrt{T_i}} \sqrt{e(\log p) \left( C_1^2 + \sigma^2(R_i)^2 \right) + \omega_i^2 \sigma^2(R_i)},$$

where the constant  $C_1$  is as in Assumption 4. Substitute  $R_i = R_1/2^{i/2}$  and  $T_i \geq C s_0^2 R_i^{-2}$  to obtain that

$$\lambda_{K_{T_0}} \leq C \frac{R_1}{s_0 2^{K_{T_0}/2}}.$$

To further bound  $\lambda_{K_{T_0}}$ , setting  $T_i$  as in eq.(32) in Agarwal et al. (2012), we have that

$$T_0 = \sum_{i=1}^{K_{T_0}} T_i \geq s_0^2 (\log d) 2^{K_{T_0}},$$

yielding that

$$\lambda_{K_{T_0}} \leq \frac{CR_1}{s_0 2^{K_{T_0}/2}} \leq C \sqrt{\frac{\log p}{T_0}}.$$

Then combining (B.8) finishes the proof.  $\square$

**Lemma 7.** Denote  $\sigma_i^2 = \sigma^2(R_i)$  and  $\|\theta - \theta_0\|_1 \leq R_i$ . Then the following results hold.

(a) With step size  $\alpha^t = \alpha/\sqrt{t}$ , we have that for any  $\omega > 0$ ,

$$\sum_{t=1}^T \alpha^{t-1} \|e_t(\theta)\|_\infty^2 \leq \sigma_i^2 \alpha \sqrt{T} + \omega \sigma_i^2 \alpha \sqrt{\log T}$$

holds with probability at least  $1 - 1/\omega^2$ ;

(b) Denote  $\theta_t$  as the solution in the  $t$ -th iteration. We have that any  $\omega > 0$ ,

$$\sum_{t=1}^T \langle e_t(\theta), \theta_t - \hat{\theta}_i \rangle \leq \omega \sigma_i R_i \sqrt{T}$$

holds with probability at least  $1 - 1/4\omega^2$ .

*Proof of Lemma 7.* To establish (a), we have that for any  $w > 0$ ,

$$\begin{aligned} & \mathbb{P} \left( \sum_{t=1}^T \alpha^{t-1} \|e_t(\theta)\|_\infty^2 > \sigma_i^2 \alpha \sqrt{T} + \omega \sigma_i^2 \alpha \sqrt{\log T} \right) \\ & \leq \mathbb{P} \left( \sum_{t=1}^T \alpha^{t-1} \|e_t(\theta)\|_\infty^2 > \omega \sigma_i^2 \alpha \sqrt{\log T} \right) \\ & \stackrel{(1)}{\leq} \frac{\mathbb{E} \left[ \left( \sum_{t=1}^T \alpha^{t-1} \|e_t(\theta)\|_\infty^2 \right)^2 \right]}{\omega^2 \sigma_i^4 \alpha^2 \log T} \stackrel{(2)}{\leq} \frac{\sum_{t=1}^T \alpha^{2(t-1)} \mathbb{E} \left[ \|e^t\|_\infty^4 \right]}{\omega^2 \sigma_i^4 \alpha^2 \log T} \\ & \stackrel{(3)}{\leq} \frac{\mathbb{E} \left[ \|e_t(\theta)\|_\infty^4 \right]}{\omega^2 \sigma_i^4} \stackrel{(4)}{\leq} \frac{1}{\omega^2}, \end{aligned}$$

where ineq.(1) is by Chebyshev's inequality, ineq.(2) is by Cauchy Schwartz inequality, ineq.(3) is by the fact that  $\sum_{t=1}^T \alpha^{2t} = \sum_{t=1}^T \alpha^2/t = \alpha^2 \log T$ , and ineq.(4) is by (B.7).

Part (b) is similarly established as follows

$$\begin{aligned}
& \mathbb{P} \left( \sum_{t=1}^T \left\langle e_t(\theta), \theta^t - \hat{\theta}_i \right\rangle > 2\omega R_i \sigma_i \sqrt{T} \right) \\
& \leq \frac{\sum_{t=1}^T \mathbb{E} \left[ \left\langle e_t(\theta), \theta^t - \hat{\theta}_i \right\rangle^2 \right]}{4\omega^2 R_i^2 \sigma_i^2 T} \\
& \leq \frac{\mathbb{E} \left[ \|e_t(\theta)\|_\infty^2 \right] \left\| \theta^t - \hat{\theta}_i \right\|_1^2}{4\omega^2 R_i^2 \sigma_i^2} \\
& \leq \frac{1}{4\omega^2}.
\end{aligned}$$

□

## C Proofs of Corollaries

### C.1 Proof of Corollary 1

*Proof of Corollary 1.* Theorem 1 and the continuous mapping theorem imply that, for any  $x$  satisfying that  $\|x\|_2 < \infty$ ,

$$\left| \exp(-x^\top \hat{\theta}) - \exp(-x^\top \theta_0) \right| \lesssim \sqrt{\frac{s_0(\log p)}{n_0}}. \quad (\text{C.1})$$

Conditional on  $X = x$ , we have

$$\begin{aligned}
& \left| \frac{\hat{Q}_{Y|X=x, Y>w_n}(\tau)}{Q_{Y|X=x, Y>w_n}(\tau)} - 1 \right| \\
& \stackrel{(1)}{=} \left| (1-\tau)^{-\exp(-x^\top \hat{\theta}) + \exp(-x^\top \theta_0)} - 1 \right| \\
& \stackrel{(2)}{\leq} (1-\tau)^{-\epsilon} |\log(1-\tau)| \left| \exp(-x^\top \hat{\theta}) - \exp(-x^\top \theta_0) \right| \\
& \stackrel{(3)}{\lesssim} \sqrt{\frac{s_0(\log p)}{n_0}},
\end{aligned}$$

where (1) is by the definition of  $\hat{Q}_{Y|X=x, Y>w_n}(\tau)$  and Assumption 1; (2) is by the mean value expansion with  $\epsilon$  between  $x^\top \hat{\theta}$  and  $x^\top \theta_0$ ; and (3) is by (C.1) and the fact that  $\tau \in (0, 1)$ . □

### C.2 Proof of Corollary 2

*Proof of Corollary 2.* Let

$$V = \mathbb{E} [\log(Y_i/w_n) \exp(X_i^\top \theta_0) X_i X_i^\top].$$

Following the proof of Theorems 3, it can be similarly shown that

$$\sqrt{n}x^\top \left(\tilde{\theta} - \theta_0\right) = \sqrt{n}\hat{u}^\top \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i + o_p(1) \quad (\text{C.2})$$

and  $x^\top (\hat{V}_3 - V_3)x \rightarrow 0$  as  $n_0, p \rightarrow \infty$ . Since  $\|x\|_2 < C$ ,  $\hat{u}^\top \{\exp(X_i^\top \theta_0) \log(Y_i/w_n) - 1\} X_i$  are sub-Gaussian, which implies that  $\sqrt{n}x^\top (\tilde{\theta} - \theta_0)(x^\top Vx)^{-1/2} \rightarrow N(0, 1)$  in distribution as  $n_0, p \rightarrow \infty$ . For the extreme quantile parameter  $q(x^\top \theta_0)$ , by the mean value expansion, we have

$$q(x^\top \tilde{\theta}) = q(x^\top \theta_0) + \dot{q}(x^\top \theta_0)x^\top \left(\tilde{\theta} - \theta_0\right) + \ddot{q}(x^\top \theta_*) \left\{x^\top \left(\tilde{\theta} - \theta_0\right)\right\}^2 / 2,$$

where  $\theta_*$  is between  $\theta_0$  and the debiased estimator  $\tilde{\theta}$ . Since  $x^\top (\tilde{\theta} - \theta_0) = O_p(n^{-1/2})$  from (C.2), we have

$$\sqrt{n} \left\{q(x^\top \tilde{\theta}) - q(x^\top \theta_0)\right\} = \sqrt{n}\dot{q}(x^\top \theta_0)x^\top \left(\tilde{\theta} - \theta_0\right) + o_p(1).$$

Note that Theorem 1 implies that  $\dot{q}(x^\top \hat{\theta}) \xrightarrow{p} \dot{q}(x^\top \theta_0)$  due to  $\|\hat{\theta} - \theta_0\|_1 \xrightarrow{p} 0$ . Hence,  $\dot{q}(x^\top \hat{\theta})^2 x^\top \hat{V}_3 x \xrightarrow{p} \dot{q}(x^\top \theta_0)^2 x^\top Vx$ . The asymptotic normality in Corollary 2 follows.  $\square$

## References

- AGARWAL, A., S. NEGAHBAN, AND M. J. WAINWRIGHT (2012): “Stochastic optimization and sparse statistical recovery: Optimal algorithms for high dimensions,” *Advances in Neural Information Processing Systems*, 25.
- BABII, A., R. T. BALL, E. GHYSELS, AND J. STRIAUKAS (2023): “Machine learning panel data regressions with heavy-tailed dependent data: Theory and application,” *Journal of Econometrics*, 237, 105315.
- BELLONI, A., V. CHERNOZHUKOV, D. CHETVERIKOV, AND Y. WEI (2018): “Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework,” *Annals of statistics*, 46, 3643.
- CAI, T. T., Z. GUO, AND R. MA (2023): “Statistical inference for high-dimensional generalized linear models with binary outcomes,” *Journal of the American Statistical Association*, 118, 1319–1332.

CHEN, X., J. D. LEE, X. T. TONG, AND Y. ZHANG (2020): “Statistical inference for model parameters in stochastic gradient descent,” *Annals of Statistics*, 48, 251–273.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWHEY, AND J. ROBINS (2018): “Double/debiased machine learning for treatment and structural parameters,” *Econometrics Journal*, 21, 1–68.

DAOUIA, A., L. GARDES, AND S. GIRARD (2013): “On kernel smoothing for extremal quantile regression,” *Bernoulli*, 19, 2557–2589.

DAOUIA, A., L. GARDES, S. GIRARD, AND A. LEKINA (2010): “Kernel estimators of extreme level curves,” *Test*, 20, 311–333.

DE HAAN, L. AND A. FERREIRA (2007): *Extreme Value Theory: An Introduction*, Springer Science and Business Media, New York.

EFROMOVICH, S. (2010): “Dimension reduction and adaptation in conditional density estimation,” *Journal of the American Statistical Association*, 105, 761–774.

FAGAN, S. AND R. GENÇAY (2010): “An introduction to textual econometrics,” *Handbook of empirical economics and finance*, 133–153.

FAN, J., W. WANG, AND Z. ZHU (2021): “A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery,” *Annals of statistics*, 49, 1239.

GABAIX, X. (2009): “Power Laws in Economics and Finance,” *Annual Review of Economics*, 1, 255–293.

——— (2016): “Power Laws in Economics: An Introduction,” *Journal of Economic Perspectives*, 30, 185–206.

GARDES, L. AND S. GIRARD (2010): “Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rain fall return levels,” *Extremes*, 13, 177–204.

GARDES, L., A. GUILLOU, AND A. SCHORGEN (2012): “Estimating the conditional tail index by integrating a kernel conditional quantile estimator,” *Journal of Statistical Planning and Inference*, 142, 1586–1598.

HALL, P. (1982): “On Some Simple Estimates of an Exponent of Regular Variation,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 37–42.

HAN, R., L. LUO, Y. LIN, AND J. HUANG (2024): “Online inference with debiased stochastic gradient descent,” *Biometrika*, 111, 93–108.

HUANG, J. AND C.-H. ZHANG (2012): “Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications,” *Journal of Machine Learning Research*, 13, 1839–1864.

IZBICKI, R. AND A. B. LEE (2016): “Nonparametric conditional density estimation in a high-dimensional regression setting,” *Journal of Computational and Graphical Statistics*, 25.

——— (2017): “Converting high-dimensional regression to high-dimensional conditional density estimation,” *Electronic Journal of Statistics*, 11, 2800–2831.

JAVANMARD, A. AND A. MONTANARI (2014): “Confidence intervals and hypothesis testing for high-dimensional regression,” *Journal of Machine Learning Research*, 15, 2869–2909.

LI, R., C. LENG, AND J. YOU (2020): “Semiparametric Tail Index Regression,” *Journal of Business & Economic Statistics*, 40, 82–95.

NEGAHBAN, S., B. YU, M. J. WAINWRIGHT, AND P. RAVIKUMAR (2009): “A unified framework for high-dimensional analysis of  $m$ -estimators with decomposable regularizers,” *Advances in neural information processing systems*, 22.

NICOLAU, J., P. M. RODRIGUES, AND M. Z. STOYKOV (2023): “Tail index estimation in the presence of covariates: Stock returns’ tail risk dynamics,” *Journal of Econometrics*, 235, 2266–2284.

RESNICK, S. (2007): *Heavy-tail phenomena: probabilistic and statistical modeling*, Springer Science & Business Media.

- TADDY, M. (2013): “Multinomial inverse regression for text analysis,” *Journal of the American Statistical Association*, 108, 755–770.
- VAN DE GEER, S., P. BÜHLMANN, Y. RITOV, AND R. DEZEURE (2014): “On asymptotically optimal confidence regions and tests for high-dimensional models,” *Annals of Statistics*, 42, 1166 – 1202.
- VAN DE GEER, S. A. (2008): “High-dimensional generalized linear models and the lasso,” *Annals of Statistics*, 36, 614.
- VERSHYNIN, R. (2010): “vershynin2010introduction,” *arXiv preprint arXiv:1011.3027*.
- WANG, H. AND C.-L. TSAI (2009): “Tail index regression,” *Journal of the American Statistical Association*, 104, 1233–1240.
- WANG, H. J. AND D. LI (2013): “Estimation of Extreme Conditional Quantiles Through Power Transformation,” *Journal of the American Statistical Association*, 108, 1062–1074.
- WANG, H. J., D. LI, AND X. HE (2012): “Estimation of High Conditional Quantiles for Heavy-Tailed Distributions,” *Journal of the American Statistical Association*, 107, 1453–1464.
- WONG, K. C., Z. LI, AND A. TEWARI (2020): “Lasso guarantees for  $\beta$ -mixing heavy-tailed time series,” *Annals of Statistics*, 48, 1124–1142.
- ZHANG, C.-H. AND S. S. ZHANG (2014): “Confidence intervals for low dimensional parameters in high dimensional linear models,” *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 217–242.
- ZHU, Z. AND W. ZHOU (2021): “Taming heavy-tailed features by shrinkage,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 3268–3276.