Adaptive and minimax optimal estimation of the tail coefficient

Alexandra Carpentier and Arlene K. H. Kim

University of Cambridge

Abstract: We consider the problem of estimating the tail index α of a distribution satisfying a (α, β) second-order Pareto-type condition, where β is the second-order coefficient. When β is available, it was previously proved that α can be estimated with the optimal rate $n^{-\frac{\beta}{2\beta+1}}$. On the contrary, when β is not available, estimating α with the optimal rate is challenging; so additional assumptions that imply the estimability of β are usually made. In this paper, we propose an adaptive estimator of α, and show that this estimator attains the rate $\big(n/\log\log n\big)^{-\frac{\beta}{2\beta+1}}$ without a priori knowledge of β and any additional assumptions. Moreover, we prove that this $\big(\log\log n\big)^{\frac{\beta}{2\beta+1}}$ factor is unavoidable by obtaining the companion lower bound.

Key words and phrases: Adaptive estimation, minimax optimal bounds, extreme value index, Pareto-type distributions.

1. Introduction

We consider the problem of estimating the tail index α of an (α, β) second-order Pareto distribution F, given n i.i.d. observations $X_1, \ldots, X_n$. More precisely, we assume that for some $\alpha, \beta, C, C' > 0$,

$$\big|1 - F(x) - Cx^{-\alpha}\big| \leq C' x^{-\alpha(1+\beta)}. \tag{1.1}$$

We will write $\mathcal{S}(\alpha, \beta) := \mathcal{S}(\alpha, \beta, C, C')$ for the set of distributions that satisfy this property (see Definition (2)). Here the tail index α characterizes the heaviness of the tail, and β represents the proximity between F and an α-Pareto distribution $F_\alpha^P : x \in [C^{1/\alpha}, \infty) \to 1 - Cx^{-\alpha}$.

There is an abundant literature on the problem of estimating α. A very popular estimator is Hill's estimator (Hill, 1975) (see also Pickands' estimator (Pickands, 1975)). Hill (1975) considered α-Pareto distribution for the tail, and suggested an estimator $\hat{\alpha}_H(r)$ of the tail index α based on the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$ where r is the fraction of order statistics from the tail,

$$\hat{\alpha}_H(r) = \left(\frac{1}{\lfloor rn \rfloor} \sum_{i=1}^{\lfloor rn \rfloor} \frac{\log(X_{(n-i+1)})}{\log(X_{(n-\lfloor rn \rfloor+1)})} \right)^{-1}. \tag{1.2}$$

For more details, see e.g. de Haan and Ferreira (2006).

Limiting distribution of Hill's estimator was first proved by Hall (1982) when β is known. Under a model that is quite similar to (1.1), he proved that if $rn^{1/(2\beta+1)} \to 0$ as $n \to \infty$, $\sqrt{nr}(\hat{\alpha}_H(r) - \alpha)$

converges in distribution to $N(0,\alpha^2)$. He also considered more restricted condition, say, the exact Hall condition,

$$\left|1 - F(x) - Cx^{-\alpha}\right| = C'' x^{-\alpha(1+\beta)} + o(x^{-\alpha(1+\beta)}). \tag{1.3}$$

Under the model (1.3) with the choice of the sample fraction $r^* = Cn^{-\frac{1}{2\beta+1}}$ with some constant C, Theorem 2 of Hall (1982) states that $n^{\beta/(2\beta+1)}(\hat{\alpha}_H(r^*) - \alpha)$ converges to a Gaussian distribution with finite mean and variance, depending on the parameters of the true distribution.

The companion lower bound $n^{-\beta/(2\beta+1)}$ under the assumption (1.1) was proved by Hall and Welsh (1984). Drees (2001) improved this result by obtaining sharp asymptotic minimax bounds again when β is available. From these results, we know that the second-order parameter β is crucial to understand the behaviour of the distribution. Indeed, it determines the rate of estimation of α as well as the optimal sample fraction.

However, β is unknown in general. To cope with this problem, Hall and Welsh (1985) proved that under condition (1.3), it is possible to estimate β in a consistent way, and thus also to estimate the sample fraction r^* consistently by $\hat{r}$ (see Theorem 4.2 in their paper). Theorem 4.1 of Hall and Welsh (1985) deduces from these results that the estimate $\hat{\alpha}_H(\hat{r})$ is asymptotically as efficient as $\hat{\alpha}_H(r^*)$, that is, $n^{\beta/(2\beta+1)}(\hat{\alpha}_H(\hat{r}) - \alpha)$ converges to a Gaussian distribution with the same mean and variance as the one resulting from the choice r^*. Their result is pointwise, but not uniform under the model (1.3), as opposed to the uniform convergence when β is known.

This first result on adaptive estimation was extended in several ways. For instance, Gomes, et. al. (2008) provided more precise ways to reduce the bias of the estimate of α using the estimate of β by supposing the third order condition. The adaptive estimates of α under the third order condition was considered in Gomes, et. al. (2012). In addition, several other methods for estimating r^* have been proposed, e.g. bootstrap (e.g. Danielsson, et. al. (2001)) or regression (e.g. Beirlant, et. al. (1996)). In particular, Drees and Kaufmann (1998) considered a method that is related to Lepski's method (see Lepski (1992) for more details in a functional estimation setting) by choosing the sample fraction that balances the squared bias and the variance of the resulting estimate. They proved that Hill's estimate computed with this sample fraction is asymptotically as efficient as the oracle estimate if F satisfies a condition that is slightly more restrictive than the condition (1.3). Finally, Grama and Spokoiny (2008) consider a more general setting than (1.1). However, when they apply their results to the exact Hall model (without little o), their estimator obtains the optimal rate up to a $\log(n)$ factor, which is clearly sub-optimal as proven in Hall and Welsh (1985).

In this paper, we focus on deriving results for the setting (1.1). Indeed, many common distributions (in particular some distributions with change points in the tail) belong to it, and the construction of the lower bound in Hall and Welsh (1984) was proved in this model. However, to the best of our knowledge, either the existing results that we mentioned previously hold in a more restrictive setting than the model (1.1), typically in a model that is close to the model (1.3) (see e.g. Hall and Welsh (1985); Beirlant, et. al. (1996); Drees and Kaufmann (1998); Danielsson, et. al. (2001); Gomes, et. al. (2008, 2012)), or the convergence rates for the setting (1.1) in the previous results are worse than one could expect (see e.g. Grama and Spokoiny (2008)). It is important to note here that the set of distributions described in Equation (1.1) is significantly larger than the set of distributions that satisfy the restricted condition (1.3). As will be explained later, the adaptive estimation in our setting (i.e. condition (1.1)) is more involved since the second-order parameter β is not always estimable (even a consistent estimator does not exist for all distributions in this model), and the adaptive procedures based on estimating β or the oracle sample fraction r^* as in the papers (Hall and Welsh (1985); Gomes, et. al. (2008, 2012)) might not work on all the functions satisfying (1.1).

The contributions of this paper are the following. We construct an adaptive estimator $\hat{\alpha}$ of α in the setting (1.1) and prove that $\hat{\alpha}$ converges to α with the rate $(n/\log\log(n))^{-\beta/(2\beta+1)}$. More precisely, for an arbitrarily small $\epsilon > 0$, and some arbitrarily large range I_1 for α and $[\beta_1, \infty)$ for β, there exist large constants $D, E > 0$ such that for any $n > D\log(\log(n)/\epsilon)$

$$\sup_{\alpha\in I_1, \beta>\beta_1} \sup_{F\in\mathcal{S}(\alpha,\beta)} \mathbb{P}_F\left(|\hat{\alpha}-\alpha| \geq E\left(\frac{n}{\log(\log(n)/\epsilon)}\right)^{-\frac{\beta}{2\beta+1}}\right) \leq \epsilon. \tag{1.4}$$

There is an additional $\big(\log\log(n)\big)^{\frac{\beta}{2\beta+1}}$ factor in the rate with respect to the oracle rate, which comes from the fact that we adapt over β on a set of distributions where β is not estimable. Although we obtain worse rates of convergence than the oracle rate, we actually prove the optimality of our adaptive estimator by obtaining a matching lower bound. Indeed, there exists a small enough constant $E' > 0$ such that for any n large enough, and for any estimator $\tilde{\alpha}$,

$$\sup_{\alpha\in I_1, \beta>\beta_1} \sup_{F\in\mathcal{S}(\alpha,\beta)} \mathbb{P}_F\left(|\tilde{\alpha}-\alpha| \geq E'\left(\frac{n}{\log(\log(n))}\right)^{-\frac{\beta}{2\beta+1}}\right) \geq \frac{1}{4}.$$

Both lower and upper bounds containing the $(\log\log(n))^{\beta/(2\beta+1)}$ factor are new to the best of our knowledge (we do not provide a tight scaling factor as in the paper by Novak (2013), but the setting in this paper is different and their rate does not involve this additional $(\log\log(n))^{\beta/(2\beta+1)}$ factor). The presence of the $\log\log n$ factor is not unusual in adaptive estimation (see Spokoiny (1996) in a signal detection setting). This issue is also discussed in the paper (Drees and Kaufmann, 1998).

The adaptive estimator $\hat{\alpha}$ we propose in this paper is based on a sequence of estimates $\hat{\alpha}(k)$ defined in (3.1), where the parameter $k \in \mathbb{N}$ plays a role similar to the sample fraction in Hill's estimator (see Subsection 3.1 for more details). These estimates $\hat{\alpha}(k)$ are not based on order statistics, but on probabilities of tail events. We first prove that for an appropriate choice of this threshold k (independent of α or β), $\hat{\alpha}(k)$ is consistent. We then prove that for an oracle choice of k (as a function of β), this estimate is minimax-optimal for distributions satisfying (1.1) with the rate $n^{-\frac{\beta}{2\beta+1}}$. Finally an adaptive version of this estimate, where the parameter k is chosen in a data-driven way without knowing β in advance, is proved to satisfy Equation (1.4).

2. Definitions of distribution classes

In this section, we introduce two sets of distributions of interest, namely the class of approximately α-Pareto distributions, and the class of approximately (α, β) second-order Pareto distributions. We let $\mathcal{D}$ be the class of distribution functions on $[0, \infty)$.

Definition 1. Let $\alpha > 0$, $C > 0$. We denote by $\mathcal{A}(\alpha, C)$ the class of approximately α-Pareto distributions:

$$\mathcal{A}(\alpha, C) = \Big\{ F \in \mathcal{D} : \lim_{x \to \infty} (1 - F(x)) x^{\alpha} = C \Big\}.$$

Distributions in $\mathcal{A}(\alpha, C)$ converge to Pareto distributions for large x, and these distributions have been used as a first attempt to understand heavy tail behavior (see Hill (1975); de Haan and Ferreira (2006)). The first-order parameter α characterizes the tail behavior such that distributions with smaller α correspond to heavier tails.

In order to provide rates of convergence (of an estimator of α), we define the set of second-order Pareto distributions.

Definition 2. Let $\alpha > 0$, $C > 0$, $\beta > 0$ and $C' > 0$. We denote by $\mathcal{S}(\alpha, \beta, C, C')$ the class of approximately (α, β) second-order Pareto distributions:

$$\mathcal{S}(\alpha, \beta, C, C') = \Big\{ F \in \mathcal{D} : \forall x \text{ s.t. } F(x) \in (0, 1], \big|1 - F(x) - Cx^{-\alpha}\big| \leq C' x^{-\alpha(1+\beta)} \Big\}. \tag{2.1}$$

From Definition 2, we know that not only are the distributions in $\mathcal{S}(\alpha, \beta, C, C')$ approximately α-Pareto, but we additionally have a bound on the rate at which they approximate Pareto distributions. This rate of approximation is linked to the second-order parameter β—a large β corresponds to a distribution that is very close to a Pareto distribution (in particular, when $\beta = \infty$, it becomes exactly Pareto), and a small β corresponds to a distribution that is well approximated by a Pareto distribution only for a very large x. From now, if there is no confusion, we call the distributions in

$\mathcal{S}(\alpha, \beta, C, C')$ second-order Pareto distributions, and we use the notation $\mathcal{A}$ and $\mathcal{S}$ without writing parameters explicitly.

The condition in (2.1) is related to the condition (1.3), but is weaker. Indeed, the condition (1.3) implies

$$\lim_{x\to\infty} \frac{1 - F(x) - Cx^{-\alpha}}{x^{-\alpha(1+\beta)}} = C',$$

whereas our condition imposes only an upper bound,

$$\limsup_{x\to\infty} \left| \frac{1 - F(x) - Cx^{-\alpha}}{x^{-\alpha(1+\beta)}} \right| \leq C'.$$

This difference is essential in the estimation problem. For instance, in the setting (1.3), it is possible to estimate β consistently (see e.g. Hall and Welsh (1985)), whereas in our setting (2.1), it is not possible to estimate β consistently over the set $\mathcal{S}$ of distributions for $\beta \in [\beta_1, \beta_2]$ with $0 < \beta_1 < \beta_2$. Adaptive estimation of α is thus likely to be more involved in our setting than in the more restricted model (1.3). For instance, many adaptive techniques rely on estimating β or the sample fraction as a function of β, which is not directly applicable in our setting (see e.g. Hall and Welsh (1985); Danielsson, et. al. (2001); Gomes, et. al. (2012)).

Remark 1. *The difference between the functions satisfying the condition in Definition 2 and the condition* (1.3) *is related to the difference between Hölder functions that actually attain their Hölder exponent and Hölder functions that are in a given Hölder ball but do not attain their Hölder exponent (see e.g. Giné and Nickl (2010) for a comparison of these two sets, and the problem for estimation when the second set is considered).*

3. Main results

Most estimates in the literature are based on order statistics (as Hill's estimate or Pickands' estimate), which causes a difficulty for one to analyse them in a non-asymptotic way. In contrast, the estimate we will present in Section 3.1 verifies large deviation inequalities in a simple way. This estimate is based on probabilities of well chosen tail events.

3.1. A new estimate

Let $X_1, \ldots, X_n$ be an i.i.d. random sample from a distribution $F \in \mathcal{A}$. We write, for any $k \in \mathbb{N}$,

$$p_k := \mathbb{P}(X > e^k) = 1 - F(e^k),$$

and its empirical estimate

$$\hat{p}_k := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i > e^k\}.$$

We define the following estimate of α for any $k \in \mathbb{N}$

$$\hat{\alpha}(k) := \log(\hat{p}_k) - \log(\hat{p}_{k+1}). \tag{3.1}$$

This estimate gives the following large deviation inequalities, which is crucial for proving consistency and convergence rates of $\hat{\alpha}(k)$.

Lemma 1 (Large deviation inequality)**.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from F.*

A. *Suppose $F \in \mathcal{A}$ and let $\delta > 0$. For any k such that $p_{k+1} \geq \frac{16\log(2/\delta)}{n}$, with probability larger than $1 - 2\delta$,*

$$\left|\hat{\alpha}(k) - (\log(p_k) - \log(p_{k+1}))\right| \leq 6\sqrt{\frac{\log(2/\delta)}{np_{k+1}}}. \tag{3.2}$$

B. *Assume now that $F \in \mathcal{S}$ and let $\delta > 0$. For any k such that $p_{k+1} \geq \frac{16\log(2/\delta)}{n}$ and $e^{-k\alpha\beta} \leq C/(2C')$, with probability larger than $1 - 2\delta$,*

$$\left|\hat{\alpha}(k) - \alpha\right| \leq 6\sqrt{\frac{\log(2/\delta)}{np_{k+1}}} + \frac{3C'}{C}e^{-k\alpha\beta} \tag{3.3}$$

$$\leq 6\sqrt{\frac{e^{(k+1)\alpha+1}\log(2/\delta)}{Cn}} + \frac{3C'}{C}e^{-k\alpha\beta}. \tag{3.4}$$

For this new estimate $\hat{\alpha}(k)$, k plays a similar role as the sample fraction in Hill's estimate (1.2). The bias-variance trade-off should be solved by choosing k in an appropriate way as a function of β (we will explain this more in details later). Choosing a too large k leads to using a small sample fraction, and the resulting estimate has a large variance and a small bias. On the other hand, choosing a too small k yields a large bias and a small variance for the estimate. The optimal k equalises the bias term and the standard deviation.

3.2. Rates of convergence

We first consider the set of approximately Pareto distributions, and prove that the estimate $\hat{\alpha}(k_n)$ is consistent if we choose k_n such that it diverges to ∞ but not too fast.

Theorem 1 (Consistency in $\mathcal{A}$)**.** *Let $F \in \mathcal{A}$. Let $k_n \in \mathbb{N}$ be such that $k_n \to \infty$ and $(\log(n)/n)e^{k_n\alpha} \to 0$ as $n \to \infty$. Then*

$$\hat{\alpha}(k_n) \to \alpha \quad a.s.$$

Choosing (for instance) $k_n = (\log\log(n))$ ensures almost sure convergence.

The estimate $\hat{\alpha}(\log\log(n))$ converges to α almost surely under the rather weak assumption that F belongs to $\mathcal{A}$. But on such sets, no uniform rate of convergence exists, and this is the reason why the restricted set $\mathcal{S}$ is introduced.

Let $\alpha, \beta, C, C' > 0$. Consider now the set $\mathcal{S} := \mathcal{S}(\alpha, \beta, C, C')$ of second-order Pareto distributions. We assume in a first instance that, although we do not have access to α, we know the parameter $\alpha(2\beta+1)$. It is not very realistic assumption, but we will explain soon how we can modify the estimate so that it is minimax optimal on the class of second-order Pareto distributions.

Theorem 2 (Rate of convergence when $\alpha(2\beta+1)$ is known)**.** *Let n be such that (4.7) is satisfied. Let $k_n^* = \lfloor \log(n^{\frac{1}{\alpha(2\beta+1)}}) + 1 \rfloor$. Then for any $\delta > 0$, we have*

$$\sup_{F \in \mathcal{S}} \mathbb{P}_F \left(|\hat{\alpha}(k_n^*) - \alpha| \geq \left(B_1 + \frac{3C'}{C} \right) n^{-\frac{\beta}{2\beta+1}} \right) \leq 2\delta,$$

where $B_1 = 6\sqrt{e^{2\alpha+1} \frac{\log(2/\delta)}{C}}$.

Theorem 2 states that, uniformly on the class of second-order Pareto distributions, the estimate $\hat{\alpha}(k_n^*)$ converges to α with the minimax optimal rate $n^{-\frac{\beta}{2\beta+1}}$ (see Hall and Welsh (1984) for the matching lower bound).

Remark 2. *Theorem 2 can be used to prove the convergence rate of our estimator by modifying the choice of k_n^*, when $\alpha(2\beta+1)$ is unknown but only β is known. For instance, we can plug a rough estimate $\tilde{\alpha} := \hat{\alpha}((\log\log(n))^2)$ of α into k_n^*. The idea behind this choice is that with sufficiently large n, we have with high probability,*

$$|\hat{\alpha}((\log\log(n))^2) - \alpha| = O\left(\frac{1}{\log n}\right).$$

Then $\hat{k}_n^1$ is defined as $\lfloor \log(n^{\frac{1}{\tilde{\alpha}(2\beta+1)}}) + 1 \rfloor$. Finally, the rate of convergence of $\hat{\alpha}(\hat{k}_n^1)$ can be shown as $n^{-\beta/(2\beta+1)}$ by proving $\exp(\hat{k}_n^1) = O(n^{1/(\alpha(2\beta+1))})$ with high probability.

However, the previous optimal choice of k (k_n^* or $\hat{k}_n^1$) still depends on β, which is unavailable in general. To deal with this problem, we construct an adaptive estimate of α that does not depend on β but still attains a rate that is quite close to the minimax optimal rate $n^{-\frac{\beta}{2\beta+1}}$ on the class of β second-order Pareto distributions.

The adaptive estimator is obtained by considering a kind of bias and variance trade-off based on the large deviation inequality (3.2). Suppose we know the optimal choice of k^*. Then this k^* will optimize the squared error by making bias and standard error (of the estimate with respect to its expectation) equal. Since the bias is decreasing while the standard error is increasing as k

increases, for all k' larger than this optimal k^*, the bias will be smaller than the standard error. Based on this heuristic (originally proposed by Lepski (1992)), we pick the smallest k which satisfies for all k' larger than k, the proxy for the bias is smaller than the proxy for the standard error $O(\sqrt{1/(n\hat{p}_{k'+1})})$ as in (3.2). For the proxy for the bias, we use $|\hat{\alpha}(k') - \hat{\alpha}(k)|$ by treating $\hat{\alpha}(k)$ as the true α based on the idea that $\hat{\alpha}(k)$ would be very close in terms of the rate to the true α (if k is selected in an optimal way).

More precisely, we choose k as follows, for $1/4 > \delta > 0$

$$\begin{aligned}\hat{k}_n = \inf\Big\{ & k \in \mathbb{N}: \quad \hat{p}_{k+1} > \frac{24\log(2/\delta)}{n} \quad \text{and} \\ & \forall k' > k \ \text{ s.t. } \ \hat{p}_{k'+1} > \frac{24\log(2/\delta)}{n}, \quad |\hat{\alpha}(k') - \hat{\alpha}(k)| \le A(\delta)\sqrt{\frac{1}{n\hat{p}_{k'+1}}}\Big\}, \end{aligned} \tag{3.5}$$

where $A(\delta)$ satisfies the condition (3.6) in the following theorem.

Theorem 3 (Rates of convergence with unknown β)**.** *Let $1/4 > \delta > 0$ and let n be such that (4.9) is satisfied. Consider the adaptive estimator $\hat{\alpha}(k_n)$ where k_n is chosen as described in (3.5) where $A(\delta)$ satisfies the following condition*

$$A(\delta) \ge 6\sqrt{2(C + C')\log(2/\delta)}\left(2\sqrt{\frac{e^{2\alpha+1}}{C}} + \frac{C'}{C}\right). \tag{3.6}$$

Then we have

$$\sup_{F \in \mathcal{S}} \mathbb{P}_F\left(|\hat{\alpha}(\hat{k}_n) - \alpha| \ge \left(B_2 + \frac{3C'}{C}\right)\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}}\right) \le \left(1 + \frac{1}{\alpha}\log\left(\frac{(C+C')n}{16}\right)\right)\delta.$$

where $B_2 = \left(B_1 + 2A(\delta)\sqrt{\frac{e^{2\alpha}}{C}}\right)\frac{1}{\sqrt{\log(2/\delta)}}$ and B_1 is defined in Theorem 2.

Theorem 3 holds for any (α, β) provided that n and $A(\delta)$ are larger than some constants depending on α, β, C, C', and on the probability δ. The advantage of our adaptive estimator is that since the threshold $\hat{k}_n$ is chosen adaptively to the samples, the second-order parameter β does not need to be known in the procedure in order to obtain the convergence rate of $\hat{\alpha}(\hat{k}_n)$. Theorem 3 gives immediately the following corollary.

Corollary 1. *Let $\epsilon \in (0, 1)$ and $C' > 0$ and let $0 < \alpha_1 < \alpha_2$ and $0 < C_1 < C_2$. We use $\hat{k}_n$ as in*

(3.5) where $A(\delta) = A(\delta(\epsilon)) =: A(\epsilon)$ *is chosen as in Equation* (4.19)*. If* n *satisfies (4.21), then*

$$\sup_{\substack{\alpha\in[\alpha_1,\alpha_2],\beta\in[\beta_1,\infty]\\ C\in[C_1,C_2]}} \sup_{F\in\mathcal{S}(\alpha,\beta,C,C')} \mathbb{P}_F\left(|\hat{\alpha}(\hat{k}_n)-\alpha| \geq B_3\left(\frac{n}{\log\left(\frac{2}{\epsilon}\left(1+\frac{\log((C_2+C')n)}{\alpha_1}\right)\right)}\right)^{-\frac{\beta}{2\beta+1}}\right) \leq \epsilon,$$

where B_3 *is a constant explicitly expressed in (4.20), which only depends on* $\alpha_2, C_1, C_2,$ *and* C'*.*

In other words, if we fix the range of the α and C and a lower bound on β to which we wish to adapt, we can tune the parameters of the adaptive choice of $\hat{k}_n$ so that we adapt to the maximal β such that F is β second-order Pareto. Moreover, this adaptive procedure works uniformly well over the set of second-order Pareto distributions satisfying (1.1) (for $\alpha \in [\alpha_1, \alpha_2], \beta \in [\beta_1, \infty], C \in [C_1, C_2]$), which is much larger than the class of distributions that verify the condition (1.3). Then this gives *non-asymptotic guarantees* with *explicit bounds.*

Remark 3. *The parameter* C' *plays a role in the definition of the second order Pareto class that is slightly different than the one of* C *or* α, β*. Unlike* α *or* C*,* C' *is not uniquely defined: if* $F \in \mathcal{S}(\alpha, \beta, C, \tilde{C}')$*, then* $F \in \mathcal{S}(\alpha, \beta, C, C')$ *with* $C' \geq \tilde{C}'$*. This implies in particular that the results of Corollary 1 could have been rewritten, fixing a constant* $C' > 0$ *and writing* $\tilde{C}'$ *for a constant that fits more closely* F*, by taking supremum over* $F \in \mathcal{S}(\alpha, \beta, C, \tilde{C}')$ *where* $\tilde{C}' \leq C'$*. Being non-adaptive over* $\tilde{C}'$ *and choosing a loose constant* C' *instead of* $\tilde{C}'$ *will only worsen the bound by a constant factor, unlike making a mistake on* β *which will worsen the exponent of the bound.*

It seems that we lose a $(\log\log(n))^{\frac{\beta}{2\beta+1}}$ factor with respect to the optimal rate, due to adaptivity to β. However, the lower bound below implies that this $(\log\log(n))^{\frac{\beta}{2\beta+1}}$ loss is inevitable; hence the rate provided in Theorem 3 is sharp.

Theorem 4 (Lower bound)**.** *Let* $\alpha_1, \beta_1, C_1, C_2, C' > 0$ *be such that* $C_1 \leq \exp(-\frac{1}{2\alpha_1(2\beta_1+1)})$*,* $C_2 \geq 1$ *and* $C' \geq \frac{1}{2\alpha_1\beta_1}$*. Let* n *be sufficiently large. Then for any estimate* $\tilde{\alpha}$ *of* α*,*

$$\sup_{\substack{\alpha\in[\alpha_1,2\alpha_1],\beta\in[\beta_1,\infty)\\ C\in[C_1,C_2]}} \sup_{F\in\mathcal{S}(\alpha,\beta,C,C')} \mathbb{P}_F\left(|\tilde{\alpha}-\alpha| \geq B_4\Big(\frac{n}{\log\big(\log(n)/2\big)}\Big)^{-\frac{\beta}{2\beta+1}}\right) \geq \frac{1}{4},$$

where B_4 *is a constant depending on* α_1 *and* β_1*, which is provided in (4.30).*

The lower bound result is proved with specific ranges of the parameters (e.g. restrictions on C_1, C_2, C' in the statement of Theorem 4), but it can be modified by considering different ranges (see Remark 4).

3.3. Additional remarks on our estimate

In the definition of our estimate, we use exponential spacings (i.e. we estimate the probability that the random variable is larger than e^k), but we can generalize our estimate by considering the probability of other tail events. For some parameters $u > v \geq 1$, define

$$\hat{q}_u = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i > u\}, \text{ and } \hat{q}_v = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i > v\}.$$

We define the following estimate of α as

$$\hat{\alpha}(u, v) = \frac{\log(\hat{q}_v) - \log(\hat{q}_u)}{\log(u) - \log(v)}. \tag{3.7}$$

If we fix $v \sim O(n^{1/(\alpha(2\beta+1))})$ and $u/v \sim O(1)$, then we will also obtain the oracle rate for estimating α with $\hat{\alpha}(u, v)$. However, the choice of u/v will have an impact on the constants. In practice, these parameters are important to tune well (in particular for the exact Pareto case, or for distributions satisfying Equation (1.3)). However, a precise analysis of the best choices for u and v (in terms of constants) is beyond the scope of this paper.

Another point we want to address is the relation between our estimate and usual estimates based on order statistics. To estimate the tail index α, it is natural to consider the quantiles associated with the tail probabilities. For the estimates based on order statistics, one fixes some tail-probabilities and then observes the order statistics in order to estimate the quantiles. On the other hand, we fix some values corresponding to the quantiles, and estimate the associated tail probabilities. Based on such a link, one could relate any existing method based on order statistics to the method based on tail probabilities.

In particular, the estimator based on order statistics corresponding to our estimator would be of the form, for some parameters $1 \geq q_v > q_u \geq 0$,

$$\tilde{\alpha}(q_u, q_v) = \frac{\log(q_v) - \log(q_u)}{\log(\hat{u}) - \log(\hat{v})}, \tag{3.8}$$

where $\hat{u} = X_{(n-\lfloor q_u n \rfloor)}$ and $\hat{v} = X_{(n-\lfloor q_v n \rfloor)}$. This estimate can be interpreted as the inverse of some generalized Pickands' estimate (see Pickands (1975), it is however *not* Pickands' estimate). There is actually a duality between these two estimators: for any couple (q_u, q_v) in the definition (3.8), it is possible to find (u, v) in the definition (3.7) such that these two estimates exactly match (see Figure 3.1 for an illustration). However, there is no analytical transformation from one estimate to the other since such a transformation will be data dependent.

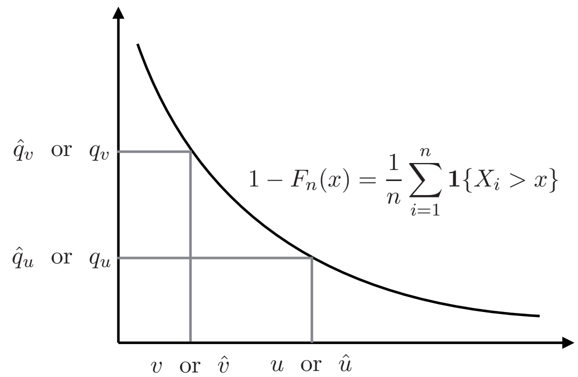

FIG 3.1. *Duality between the estimate* (3.7) *and the estimate* (3.8).

Acknowledgment

We are grateful to Richard J. Samworth and Richard Nickl for their comments and advice. We are also grateful to the anonymous Referees, the Associate Editor and the Editor for very insightful comments that were helpful in enhancing the quality of the paper.

References

Beirlant, J., and Vynckier, P., and Teugels, J. (1996). Tail index estimation, Pareto quantile plots and regression. *Journal of American Statistical Association*, **70**, 1659–1667.

Cover, T. M. and Thomas, J. A. (2012). *Elements of Information Theory.* Wiley-interscience.

Danielsson, J. and de Haan, L. and Peng, L. and de Vries, C.G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *Journal of Multivariate Analysis*, **2**, 226–248

Drees, H. (2001). Minimax risk bounds in extreme value theory. *The Annals of Statistics*, **29**, (1) 266–294.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and Their Applications*, **75**, 149–172.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics*, **38**, (2) 1122–1170.

Gomes, M. I. and F. Figueiredo, and Neves, M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action. *Extremes*, **15**. 463–489

Gomes, I. M and De Haan, L., and Rodrigues, Lígia Henriques (2008). Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society: Series B*, **91**, 31–52.

Grama, I. and Spokoiny, V. (2008). Statistics of extremes by oracle estimation. *The Annals of Statistics*, **36**, (4) 1619–1648.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer series in operations research und financial engineering.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society: Series B.*, **44**, (1) 37–42.

Hall. P. and Welsh, A. H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *The Annals of Statistics*, **12**, (3) 1079–1084.

Hall. P. and Welsh, A. H. (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, **75**, (1) 331–341.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, (5) 1163–1174.

Lepski, O. V. (1992). On problems of adaptive estimation in white gaussian noise. *Topics in nonparametric estimation*, **12**, 87–106.

Novak, S. Y. (2013). Lower bounds to the accuracy of inference on heavy tails. *Bernoulli*(to appear)

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 119–131.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, **24**, (6) 2477–2498.

Tsybakov, A. B.(2008). *Introduction to Nonparametric Estimation.* Springer.

Van der Vaart, A. W.(2000). *Asymptotic Statistics.* Cambridge University Press.

University of Cambridge

E-mail: a.carpentier@statslab.cam.ac.uk

University of Cambridge

E-mail: a.kim@statslab.cam.ac.uk

Supplementary Material for the paper : "Adaptive and minimax optimal estimation of the tail coefficient"

Alexandra Carpentier and Arlene K.H. Kim

4. Technical proofs

Lemma 2 contains a classical and simple, yet important result for the paper.

Lemma 2 (Bernstein inequality for Bernoulli random variables)**.** *Let $X_1, \ldots, X_n$ be i.i.d. observations from F, and we define $p_k = 1 - F(e^k)$ and $\hat{p}_k = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i > e^k\}$. Let $\delta > 0$ and also let n be large enough so that $p_k \geq \frac{4\log(2/\delta)}{n}$. Then with probability $1-\delta$,*

$$|\hat{p}_k - p_k| \leq 2\sqrt{\frac{p_k \log(2/\delta)}{n}}. \tag{4.1}$$

Proof of Lemma 2. The proof is using Bernstein inequality (e.g. see Lemma 19.32 of Van der Vaart (2000)) of the following form; for any bounded, measurable function g, we have for every $t > 0$,

$$\mathbb{P}\left(\left|\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}g(X)\right)\right| > t\right) \leq 2\exp\left(-\frac{1}{4}\frac{t^2}{\mathbb{E}g^2 + t||g||_\infty/\sqrt{n}}\right).$$

We use $g(\cdot) = \mathbf{1}\{\cdot > e^k\}$ and $t = 2\sqrt{p_k\log(2/\delta)}$ in the above inequality. Using the fact that $t = 2\sqrt{p_k\log(2/\delta)} \leq \sqrt{n}p_k$ by the assumption of $p_k \geq (4\log(2/\delta))/n$, we have

$$\begin{aligned}\mathbb{P}\left(\sqrt{n}|\hat{p}_k - p_k| > t\right) &\leq 2\exp\left(-\frac{1}{4}\frac{t^2}{p_k + t/\sqrt{n}}\right)\\ &\leq 2\max\left[\exp\left(-\frac{1}{4}\frac{t^2}{p_k}\right), \exp\left(-\frac{1}{4}\sqrt{n}t\right)\right]\\ &\leq 2\exp\left(-\frac{1}{4}\frac{t^2}{p_k}\right)\\ &= \delta,\end{aligned}$$

where the last equality follows by definition of t. □

Proof of Lemma 1. **A.** Since $p_k \geq 16\log(2/\delta)/n$, we can use Lemma 2. Rewriting the inequality (4.1), we have with probability larger than $1-\delta$

$$\log\left(1 - 2\sqrt{\frac{\log(2/\delta)}{np_k}}\right) \leq \log(\hat{p}_k) - \log(p_k) \leq \log\left(1 + 2\sqrt{\frac{\log(2/\delta)}{np_k}}\right).$$

Then using the simple inequalities $\log(1+u) \le u$, and $\log(1-u) \ge (-3u)/2$ for $u < 1/2$,

$$\log(p_k) - 3\sqrt{\frac{\log(2/\delta)}{np_k}} \le \log(\hat{p}_k) \le \log(p_k) + 2\sqrt{\frac{\log(2/\delta)}{np_k}}.$$

By using a similar inequality for $\log(\hat{p}_{k+1})$, with probability larger than $1-2\delta$,

$$\begin{aligned} \left|\hat{\alpha}(k) - (\log(p_k) - \log(p_{k+1}))\right| &\le 3\sqrt{\frac{\log(2/\delta)}{np_k}} + 3\sqrt{\frac{\log(2/\delta)}{np_{k+1}}} \\ &\le 6\sqrt{\frac{\log(2/\delta)}{np_{k+1}}}. \end{aligned} \tag{4.2}$$

B. By definition of second-order Pareto distributions, we have $\left|p_k - Ce^{-k\alpha}\right| \le C'e^{-k\alpha(1+\beta)}$, or equivalently,

$$\left|\frac{e^{k\alpha}p_k}{C} - 1\right| \le \frac{C'}{C}e^{-k\alpha\beta}.$$

Since we assume $\frac{C'}{C}e^{-k\alpha\beta} \le 1/2$, we have

$$\left|\log(p_k) - \log(C) + k\alpha\right| \le \frac{3C'}{2C}e^{-k\alpha\beta}.$$

A similar result also holds for p_{k+1}, and thus

$$\left|\log(p_k) - \log(p_{k+1}) - \alpha\right| \le \frac{3C'}{C}e^{-k\alpha\beta}. \tag{4.3}$$

Combining Equations (4.2) and (4.3), we obtain the large deviation inequality (3.3). Now, using the property of the second-order Pareto distributions, we can bound p_{k+1} from below.

$$\begin{aligned} p_{k+1} &\ge Ce^{-(k+1)\alpha}\left(1 - \frac{C'}{C}e^{-(k+1)\alpha\beta}\right) \\ &\ge \frac{C}{2}e^{-(k+1)\alpha} \ge Ce^{-(k+1)\alpha-1}, \end{aligned}$$

where the second inequality comes from the assumption that $e^{-k\alpha\beta} \le C/(2C')$. By substituting this into the inequality (3.3), the final inequality (3.4) follows. □

Proof of Theorem 1. The proof consists of the two steps—bounding the bias, and bounding the deviations of the estimate—as in the proof of the Lemma 1.B.

First, we bound the bias (more precisely, a proxy for the bias) using the property of the distribution class $\mathcal{A}$. By definition, we know that for any ϵ such that $C/2 > \epsilon > 0$, there exists a constant $B > 0$ such that for $x > B$,

$$\left|1 - F(x) - Cx^{-\alpha}\right| \le \epsilon x^{-\alpha}.$$

Since $k_n \to \infty$ as $n \to \infty$, for any n larger than some large enough N_1 (i.e. such that $\forall n \geq N_1$, $e^{k_n} > B$), we have

$$\left|p_{k_n} - Ce^{-k_n\alpha}\right| \leq \epsilon e^{-k_n\alpha}, \tag{4.4}$$

which yields since $\epsilon < C/2$, $\left|\log(p_{k_n}) - \log(C) + k_n\alpha\right| \leq \frac{3\epsilon}{2C}$ using the same technique as for the proof of Lemma 1. This holds also for $k_n + 1$ and thus

$$\left|\log(p_{k_n}) - \log(p_{k_n+1}) - \alpha\right| \leq \frac{3\epsilon}{C}. \tag{4.5}$$

Note also that Equation (4.4) can be used to bound the p_{k_n+1} below as follows.

$$p_{k_n+1} \geq (C-\epsilon)e^{-(k_n+1)\alpha} \geq \frac{C}{e^{\alpha+1}}e^{-k_n\alpha}. \tag{4.6}$$

Since $(\log(n)e^{k_n\alpha})/n \to 0$ as $n \to \infty$, we know that there exists N_2 large enough, such that for any $n \geq N_2$, $p_{k_n+1} \geq 32\log(n)/n$.

Then we can bound the proxy for the standard deviation using the result (3.2) in Lemma 1.A. For $n \geq \max(N_1, N_2)$, combining Equation (4.5) and Equation (3.2) with $\delta = 2/n^2$, we have with probability larger than $1 - 4/n^2$,

$$\left|\hat{\alpha}(k_n) - \alpha\right| \leq 6\sqrt{\frac{\log(n^2)}{np_{k_n+1}}} + \frac{3\epsilon}{C}.$$

Then we bound the first term in the right side of the above inequality using (4.6). That is,

$$6\sqrt{\frac{\log(n^2)}{np_{k_n+1}}} \leq 6\sqrt{e^{\alpha+1}\frac{\log(n^2)}{Cne^{-k_n\alpha}}} \leq \frac{6e^{(\alpha/2)+1}}{\sqrt{C}}\sqrt{\frac{\log(n)e^{k_n\alpha}}{n}}$$

By the assumption that $(\log(n)e^{k_n\alpha})/n \to 0$, and since the above inequality holds for any $\epsilon > 0$, we conclude that α_n converges in probability to α. Moreover, since $\sum_n(4/n^2) < \infty$, Borel–Cantelli Lemma says that $\hat{\alpha}(k_n)$ converges to α almost surely. □

Proof of Theorem 2. Let n satisfy the following,

$$n > \max\Big((\frac{2C'}{C})^{\frac{2\beta+1}{\beta}}, (\frac{32\log(2/\delta)e^{2\alpha}}{C})^{\frac{2\beta+1}{2\beta}}\Big). \tag{4.7}$$

We let $k^* = k_n^*$ such that $k_n^* := \left\lfloor \log(n^{\frac{1}{\alpha(2\beta+1)}}) + 1\right\rfloor$. Note that for n larger than $(2C'/C)^{\frac{2\beta+1}{\beta}}$, we have $e^{-k^*\alpha\beta} \leq C/(2C')$. This implies, together with the second-order Pareto assumption,

$$p_{k^*+1} \geq \frac{C}{2}n^{-\frac{1}{2\beta+1}}e^{-2\alpha} \geq \frac{16\log(2/\delta)}{n}$$

where the last inequality follows by assuming $n \geq (\frac{32\log(2/\delta)e^{2\alpha}}{C})^{\frac{2\beta+1}{2\beta}}$.

By (3.4) and by the choice of k_n, we have with probability larger than $1-2\delta$,

$$\left|\hat{\alpha}(k^*) - \alpha\right| \leq \left(6\sqrt{e^{2\alpha+1}\frac{\log(2/\delta)}{C}} + \frac{3C'}{C}\right) n^{-\frac{\beta}{2\beta+1}}.$$

□

The following lemma is going to be a useful tool for the proof of Theorem 3.

Lemma 3. *We define K such that $p_K \geq \frac{16\log(2/\delta)}{n}$ and also $p_{K+1} < \frac{16\log(2/\delta)}{n}$. Then for any $k \geq K+1$, with probability larger than $1-\delta$,*

$$\hat{p}_k \leq \frac{24\log(2/\delta)}{n}. \tag{4.8}$$

Proof of Lemma 3. We let $q := 16\log(2/\delta)/n$ and define a Bernoulli random variable $Y_i(q)$ (independent from $X_1, \ldots, X_n$) where $P(Y_i(q) = 1) = q$ for $i = 1, \ldots, n$. Then we compare $m_q := \frac{1}{n}\sum_{i=1}^n Y_i(q)$ and $\hat{p}_{K+1} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i > e^{K+1}\}$. Since $q > p_{K+1}$, the distribution of $\hat{p}_{K+1}$ is stochastically dominated by the distribution of m_q (that is, $P(\hat{p}_{K+1} > t) \leq P(m_q > t)$). By Lemma 2, we have with probability larger than $1-\delta$,

$$|m_q - q| \leq 2\sqrt{\frac{q\log(2/\delta)}{n}} = \frac{8\log(2/\delta)}{n}.$$

Then by stochastic dominance, with probability $1-\delta$,

$$\hat{p}_{K+1} \leq q + 2\sqrt{\frac{q\log(2/\delta)}{n}} = \frac{24\log(2/\delta)}{n}.$$

Thus, for any $k \geq K+1$ using the monotonicity of $\hat{p}_k$ (that is, $\hat{p}_k \geq \hat{p}_{k+1}$), we obtain that (4.8) holds with probability larger than $1-\delta$ as required. □

Proof of Theorem 3. The proof is based on 5 steps. We first define an event ξ in (4.11) of high probability where the deviation of empirical probabilities $\hat{p}_k$ from p_k is well upper bounded (with the same bound in the large deviation inequality in (4.1) but without a probability statement) for a given subset of indices $k \leq K$, where K is of order of $\log n$. Then we define $\bar{k}$ which is slightly smaller than the oracle k^* and also $\bar{k} \leq K$ so that on ξ the deviation of $\hat{\alpha}(\bar{k})$ from α (i.e. $|\hat{\alpha}(\bar{k}) - \alpha|$) is upper bounded as in (4.14). In the third step, we show that $\hat{p}_{\bar{k}+1} > 24\log(2/\delta)/n$ on ξ so that $\bar{k}$ is one possible index for $\hat{k}_n$. Also we prove that $\hat{k}_n \leq \bar{k}$ in Step 4 which leads us to bound $|\hat{\alpha}(\bar{k}) - \hat{\alpha}(\hat{k}_n)|$ from above on ξ using the definition of $\hat{k}_n$. This combined with the second step finally gives an upper bound of $|\hat{\alpha}(\hat{k}_n) - \alpha|$ on ξ. More precisely, we prove that on the set ξ, we have $|\hat{\alpha}(\hat{k}_n) - \alpha| \leq (B_2 + \frac{3C'}{C})(\frac{n}{\log(2/\delta)})^{-\beta/(2\beta+1)}$ where B_2 is a constant which will be defined in the last stage of the proof. Then we can bound $\mathbb{P}(|\hat{\alpha}(\hat{k}_n) - \alpha| \geq (B_2 + \frac{3C'}{C})(\frac{n}{\log(2/\delta)})^{-\beta/(2\beta+1)}) \leq \mathbb{P}(\xi^c)$ which has a small probability.

Let $F \in \mathcal{S}(\alpha, \beta, C, C')$ and $1/4 > \delta > 0$. Also we let n satisfy the following,

$$n > \log\big(\frac{2}{\delta}\big) \max\left[32\Big(\frac{2C'}{C^{1+\beta}}\Big)^{1/\beta}, \Big(\frac{32e^{2\alpha}}{C}\Big)^{\frac{2\beta+1}{2\beta}}, \Big(\frac{2C'}{C}\Big)^{\frac{2\beta+1}{\beta}}, \Big(\frac{96e^{2\alpha}}{C}\Big)^{\frac{2\beta+1}{\beta}}\right]. \tag{4.9}$$

Step 1: Definition of an event of high probability

First, we define $K \in \mathbb{N}$ such that $p_K \geq \frac{16\log(2/\delta)}{n} > p_{K+1}$. By inverting the condition for the second-order Pareto distributions, $\frac{16\log(2/\delta)}{n} \leq p_K \leq (C + C')e^{-K\alpha}$ gives $K \leq \frac{1}{\alpha}\log\left(\frac{(C+C')n}{16\log(2/\delta)}\right)$. Set $u = \frac{1}{\alpha}\log\left(\frac{Cn}{32\log(2/\delta)}\right) - 1$. Then since $n > 32(\frac{2C'}{C^{1+\beta}})^{1/\beta}\log(2/\delta)$, we know by definition of $\mathcal{S}$ that $1 - F(e^{u+1}) > \frac{16\log(2/\delta)}{n}$. Using the fact that $1 - F(e^x)$ is a decreasing function of x and $\frac{16\log(2/\delta)}{n} > p_{K+1}$, we have $u < K$. Thus we obtain the range of K by

$$\frac{1}{\alpha}\log\left(\frac{Cn}{32\log(2/\delta)}\right) - 1 < K \leq \frac{1}{\alpha}\log\left(\frac{(C+C')n}{16\log(2/\delta)}\right). \tag{4.10}$$

We define the following event

$$\xi = \Big\{\omega : \forall k \leq K, \big|\hat{p}_k(\omega) - p_k\big| \leq 2\sqrt{\frac{p_k\log(2/\delta)}{n}}, \hat{p}_{K+1}(\omega) \leq \frac{24\log(2/\delta)}{n}\Big\}. \tag{4.11}$$

By definition, we have $p_K \geq \frac{16\log(2/\delta)}{n}$, which gives the Bernstein inequality (4.1) with probability $1-\delta$ for $k \leq K$. In addition, Lemma 3 gives (4.8) with probabiltiy $1-\delta$. Thus, an union bound implies that $\mathbb{P}(\xi) \geq 1-(K+1)\delta$. By monotonicity of $\hat{p}_k$, we have on the event ξ, for any $k \geq K+1$, $\hat{p}_k \leq \frac{24\log(2/\delta)}{n}$. This implies that on the event ξ, the k, k' considered in Equation (3.5) are smaller than K and in particular, we have $\hat{k}_n \leq K$.

Step 2: Bounding the deviation of $\hat{\alpha}(k)$ from α on ξ (where $k \leq K$)

We define $\bar{k}_n = \bar{k} \in \mathbb{N}$ such that

$$\bar{k} := \left\lfloor \log\left((\frac{n}{\log(2/\delta)})^{\frac{1}{\alpha(2\beta+1)}}\right) + 1\right\rfloor.$$

By definition of $\bar{k}$, we know that $\bar{k} < K$. Indeed, by assuming $n \geq (32\frac{e^{2\alpha}}{C})^{\frac{2\beta+1}{2\beta}}\log(2/\delta)$ and by (4.10),

$$\bar{k} \leq \log\left(\Big(\frac{n}{\log(2/\delta)}\Big)^{\frac{1}{\alpha(2\beta+1)}}\right) + 1 \leq \frac{1}{\alpha}\log\left(\frac{Cn}{32\log(2/\delta)}\right) - 1 < K.$$

Thus,

$$e^{-K\alpha\beta} \leq e^{-\bar{k}\alpha\beta} \leq C/(2C'), \tag{4.12}$$

where the second inequality follows since $n > \log(2/\delta)(\frac{2C'}{C})^{\frac{2\beta+1}{\beta}}$.

Note also that $\bar{k} \leq k^*$, where $k^* := \left\lfloor \log\left(n^{\frac{1}{\alpha(2\beta+1)}}\right) + 1 \right\rfloor$ as before.

If $k < K$ satisfies $e^{-k\alpha\beta} \leq C/(2C')$, then since $p_{k+1} \geq p_K \geq (16\log(2/\delta))/n$, then using the exactly same proof as for Lemma 1.B, we have on ξ that

$$|\hat{\alpha}(k) - \alpha| \leq 6\sqrt{\frac{e^{(k+1)\alpha+1}\log(2/\delta)}{Cn}} + \frac{3C'}{C}e^{-k\alpha\beta}. \tag{4.13}$$

Since $e^{-\bar{k}\alpha\beta} \leq C/(2C')$ by (4.12) and $\bar{k} < K$, Equation (4.13) is verified for $\bar{k}$ on ξ. Then by definition of $\bar{k}$ in Equation (4.13), we have on ξ that

$$|\hat{\alpha}(\bar{k}) - \alpha| \leq \left(6\sqrt{\frac{e^{2\alpha+1}}{C}} + \frac{3C'}{C}\right)\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}}. \tag{4.14}$$

Step 3: Proof that $\hat{p}_{\bar{k}+1} > \frac{24\log(2/\delta)}{n}$ on ξ

By definition, we have on ξ, using $\bar{k} \leq K-1$ and $p_{\bar{k}+1} \geq p_K \geq (16\log(2/\delta))/n$,

$$\hat{p}_{\bar{k}+1} \geq p_{\bar{k}+1}\left(1 - 2\sqrt{\frac{\log(2/\delta)}{np_{\bar{k}+1}}}\right) \geq \frac{p_{\bar{k}+1}}{2}.$$

Then using the second order Pareto property with $(C'/C)e^{-\bar{k}\alpha\beta} \leq 1/2$, we have $p_{\bar{k}+1} \geq (Ce^{-(\bar{k}+1)\alpha})/2$, which gives

$$\hat{p}_{\bar{k}+1} \geq \frac{Ce^{-(\bar{k}+1)\alpha}}{4} \geq \frac{Ce^{-2\alpha}}{4}\left(\frac{\log(2/\delta)}{n}\right)^{1/(2\beta+1)}, \tag{4.15}$$

where the second inequality follows from $n > \log(2/\delta)(\frac{2C'}{C})^{\frac{2\beta+1}{\beta}}$ and from the definition of $\bar{k}$. Since $n > \left(\frac{96e^{2\alpha}}{C}\right)^{\frac{2\beta+1}{\beta}}\log(2/\delta)$, we have shown that $\hat{p}_{\bar{k}+1}$ is larger than $\frac{24\log(2/\delta)}{n}$ on ξ.

Step 4: Proof that $\hat{k}_n \leq \bar{k}$ on ξ

Suppose that $\hat{k}_n > \bar{k}$. Then by definition of $\hat{k}_n$ in (3.5), on ξ, there exists $k > \bar{k}$ such that $\hat{p}_{k+1} > \frac{24\log(2/\delta)}{n}$ (this imposes $k < K$ on ξ) and

$$|\hat{\alpha}(k) - \hat{\alpha}(\bar{k})| > A(\delta)\sqrt{\frac{1}{n\hat{p}_{k+1}}} \geq \frac{A(\delta)}{\sqrt{2(C+C')}}\sqrt{\frac{e^{k\alpha}}{n}}, \tag{4.16}$$

where the second inequality in the above follows by bounding $\hat{p}_{k+1}$ above by definition of ξ,

$$\hat{p}_{k+1} \leq p_{k+1}\left(1 + 2\sqrt{\frac{\log(2/\delta)}{np_{k+1}}}\right) \leq \frac{3}{2}p_{k+1} \leq 2(C+C')e^{-k\alpha},$$

where the penultimate inequality is obtained by $p_k \geq p_K \geq 16\log(2/\delta)/n$ (since $k \leq K$), and the last inequality follows by definition of the second order Pareto condition.

Since $k \geq \bar{k}+1$, we bound $e^{-k\alpha\beta} \leq e^{-\bar{k}\alpha\beta} \leq C/(2C')$ by (4.12). Also we have $p_{k+1} \geq \frac{16\log(2/\delta)}{n}$, since $p_{k+1} \geq p_K$. Equation (4.13) is thus verified on ξ for such $k > \bar{k}$. Now using $\sqrt{\frac{e^{k\alpha}\log(2/\delta)}{n}} > e^{-k\alpha\beta}$ (since $k > \bar{k}$), we have

$$|\hat{\alpha}(k) - \alpha| \leq \Big(6\sqrt{\frac{e^{\alpha+1}}{C}} + \frac{3C'}{C}\Big)\sqrt{\frac{e^{k\alpha}\log(2/\delta)}{n}}. \tag{4.17}$$

Equations (4.16) and (4.17) imply that on ξ,

$$\begin{aligned}|\hat{\alpha}(\bar{k}) - \alpha| &> \Big(\frac{A(\delta)}{\sqrt{2(C+C')}} - \sqrt{\log(2/\delta)}\big(6\sqrt{\frac{e^{\alpha+1}}{C}} + \frac{3C'}{C}\big)\Big)\sqrt{\frac{e^{k\alpha}}{n}} \\ &\geq \big(6\sqrt{\frac{e^{2\alpha+1}}{C}} + \frac{3C'}{C}\big)\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}},\end{aligned}$$

since we assume that $\frac{A(\delta)}{\sqrt{2(C+C')}} \geq 2\sqrt{\log(2/\delta)}\big(6\sqrt{\frac{e^{2\alpha+1}}{C}} + \frac{3C'}{C}\big)$. This contradicts Equation (4.14), and this means that on ξ, $\hat{k}_n \leq \bar{k}$.

Step 5: Large deviation inequality for an adaptive estimator

We have $\hat{p}_{\bar{k}+1} \geq \frac{24\log(2/\delta)}{n}$ from Step 3, and $\hat{k}_n \leq \bar{k}$ from Step 4 on ξ. Thus by definition of $\hat{k}_n$ in (3.5), we have on ξ that

$$\begin{aligned}|\hat{\alpha}(\bar{k}) - \hat{\alpha}(\hat{k}_n)| &\leq A(\delta)\sqrt{\frac{1}{n\hat{p}_{\bar{k}+1}}} \\ &\leq 2A(\delta)\sqrt{\frac{e^{2\alpha}}{C}}\left(\log\left(\frac{2}{\delta}\right)\right)^{-\frac{1}{2(2\beta+1)}} n^{-\frac{\beta}{2\beta+1}} \\ &= 2A(\delta)\sqrt{\frac{e^{2\alpha}}{C\log(2/\delta)}}\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}},\end{aligned} \tag{4.18}$$

where the second inequality follows on ξ by Equation (4.15).

Hence, Equations (4.18) and (4.14) imply that on ξ

$$|\hat{\alpha}(\hat{k}_n) - \alpha| \leq \left(\big(6\sqrt{\frac{e^{2\alpha+1}}{C}} + \frac{3C'}{C}\big) + 2A(\delta)\sqrt{\frac{e^{2\alpha}}{C\log(2/\delta)}}\right)\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}}.$$

Denote $B_1 = 6\sqrt{\frac{e^{2\alpha+1}}{C}\log(2/\delta)}$ and $B_2 = \Big(B_1 + 2A(\delta)\sqrt{\frac{e^{2\alpha}}{C}}\Big)\frac{1}{\sqrt{\log(2/\delta)}}$. Then since $\mathbb{P}(\xi) \geq$

$1-(K+1)\delta$, we have shown that

$$\sup_{F\in\mathcal{S}} \mathbb{P}_F\left(|\hat\alpha(\hat k_n)-\alpha| \geq \left(B_2+\frac{3C'}{C}\right)\left(\frac{n}{\log(2/\delta)}\right)^{-\frac{\beta}{2\beta+1}}\right)$$
$$\leq (K+1)\delta \leq \left(\frac{1}{\alpha}\log\left(\frac{(C+C')n}{16}\right)+1\right)\delta$$

where the last inequality follows by (4.10). This concludes the proof. □

Proof of Corollary 1. Set

$$\epsilon = \left(1+\frac{1}{\alpha_1}\log\big((C_2+C')n\big)\right)\delta,$$
$$A(\epsilon) = 6\sqrt{2(C_2+C')}\left(\sqrt{\log\left(\frac{2}{\epsilon}\Big(1+\frac{\log((C_2+C')n)}{\alpha_1}\Big)\right)}\Big(2\sqrt{\frac{e^{2\alpha_2+1}}{C_1}}+\frac{C'}{C_1}\Big)\right), \tag{4.19}$$

and plug $\delta=\delta(\epsilon)=\epsilon/(1+\log((C_2+C')n))/\alpha_1)$ and $A(\epsilon):=A(\delta(\epsilon))$ in the adaptive method described in Theorem 3. Set

$$B_3 := 6\sqrt{\frac{e^{2\alpha_2+1}}{C_1}}+24\frac{e^{2\alpha_2}}{C_1}\sqrt{2e(C_2+C')}+12e^{\alpha_2}\frac{C'}{C_1}\sqrt{2\frac{(C_2+C')}{C_1}}+\frac{3C'}{C_1}. \tag{4.20}$$

It holds for any $\alpha\in[\alpha_1,\alpha_2]$, $C\in[C_1,C_2]$ and $\beta>\beta_1$ that the constant in Theorem 3 can be bounded as

$$B_2+\frac{3C'}{C} = 6\sqrt{\frac{e^{2\alpha+1}}{C}}+12\sqrt{2\frac{e^{2\alpha}}{C}(C_2+C')}\Big(2\sqrt{\frac{e^{2\alpha_2+1}}{C_1}}+\frac{C'}{C_1}\Big)+\frac{3C'}{C}$$
$$\leq B_3,$$

so B_3 is a uniform bound on the constant in Theorem 3 for all considered values of α, C, β. Also, the uniform condition for the sample size is derived from Equation (4.9) by

$$n > \log\left(\frac{2}{\epsilon}\Big(1+\frac{\log((C_2+C')n)}{\alpha_1}\Big)\right)$$
$$\times\max\left[32\Big(\frac{2\bar C'}{\bar C_1^{1+\beta_1}}\Big)^{\frac{1}{\beta_1}},\Big(\frac{2\bar C'}{\bar C_1}\Big)^{2+\frac{1}{\beta_1}},\Big(\frac{32e^{2\alpha_2}}{\bar C_1}\Big)^{1+\frac{1}{2\beta_1}},\Big(\frac{96e^{2\alpha_2}}{\bar C_1}\Big)^{2+\frac{1}{\beta_1}}\right], \tag{4.21}$$

where $\bar C_1=\min(1,C_1)$ and $\bar C'=\max(1,C')$. □

Proof of Theorem 4. We prove the minimax lower bound by Fano's method (see e.g. Section 2.7 in Tsybakov (2008)). We define a set of approximately $\log(n)$ functions F_i whose first and second order parameters are respectively α_i and β_i. Until a point K_i, each distribution F_i matches a Pareto distribution with the first order parameter α, which is the same for all of the F_i. After this point K_i, F_i is Pareto with parameter α_i. These functions satisfy several specific properties summarized in Lemma 4. For instance, they are such that the for any $i \neq j$, the distance between α_i and α_j is at least of order $(\frac{n}{\log\log(n)})^{-\frac{\beta_i}{2\beta_i+1}}$. Moreover, the Kullback Leibler (KL) divergence between F_i and F_j is small enough so that F_i and F_j cannot be distinguishable as n increases. These two main properties enable us to apply Fano's lemma, which results in the lower bound of Theorem 4. For the proof, we assume that n is sufficiently large.

Step 1: Construction of a finite set of distributions

Let $\alpha > 0$ and $\beta > 1$. Let $\upsilon := \min\big(1, \frac{\alpha^2}{8\exp(\frac{1}{\alpha(2\beta-1)})}\big)$. Let $M > 1$ be an integer such that

$$\lfloor \log(n/\log(M)) \rfloor + 1 = M,$$

which implies that $\log(n)/2 < M < 2\log(n)$ for large n. Set for any integer $1 \le i \le M$

$$\begin{aligned}
\beta_i &= \beta - \frac{i}{M} \\
\gamma_i &= \frac{\beta_i}{2\beta_i+1}\Big(1 + \frac{\log(\upsilon)}{\log\log M}\Big) \\
K_i &= n^{\frac{1}{\alpha(2\beta_i+1)}}\big(\log M\big)^{-\frac{\gamma_i}{\alpha\beta_i}} = \Big(\frac{n}{\upsilon\log(M)}\Big)^{\frac{1}{\alpha(2\beta_i+1)}} \qquad (4.22)\\
t_i &= K_i^{-\alpha\beta_i} = n^{-\frac{\beta_i}{2\beta_i+1}}\big(\log M\big)^{\gamma_i} = \Big(\frac{n}{\upsilon\log(M)}\Big)^{-\frac{\beta_i}{2\beta_i+1}} \\
\alpha_i &= \alpha - t_i = \alpha - n^{-\beta_i/(2\beta_i+1)}(\log(M))^{\gamma_i}.
\end{aligned}$$

By definition, for $i < j$, we have $\beta_i > \beta_j$, $\gamma_i > \gamma_j$, $K_i < K_j$, $t_i < t_j$ and $\alpha_i > \alpha_j$. By assuming n large enough, we suppose that $\gamma_i > 0$ for all $i = 1, \dots, M$, and $\frac{\min(\alpha,1/\alpha)}{2} n^{\frac{\beta_i}{2\beta_i+1}} > M^{\frac{\beta_i}{2\beta_i+1}+1}$. Also we have $\beta_i \ge \beta - 1$, $K_i > 1$, and $\alpha - t_i \ge \alpha/2 =: \alpha_1$ for large enough n.

Using these notation, we introduce the distribution

$$1 - F_0(x) = x^{-\alpha}, \qquad (4.23)$$

and for any integer $1 \le i \le M$, we introduce *perturbed versions* of the distribution F_0

$$1 - F_i(x) = x^{-\alpha}\mathbf{1}\{1 \le x \le K_i\} + K_i^{-t_i} x^{-\alpha+t_i}\mathbf{1}\{x > K_i\}. \qquad (4.24)$$

We write $\{f_0, f_1, \dots, f_M\}$ for the densities associated with distributions $\{F_0, F_1, \dots, F_M\}$ with respect to Lebesgue measure.

Step 2: Properties of the constructed distributions

The following lemma highlights important characteristics of distributions $\{F_i, i = 1, \dots, M\}$ and their parameters corresponding to the second order Pareto distributions.

Lemma 4. *Let $1 \le i \le M$ and $1 \le j \le M$. It holds that for F_i defined as (4.24) and using notation in (4.22),*

$$F_i \in \mathcal{S}\left(\alpha - t_i, \beta_i, K_i^{-t_i}, \frac{1}{\alpha(\beta-1)}\right). \tag{4.25}$$

Moreover

$$\exp\big(-\frac{1}{\alpha(2\beta-1)}\big) \le K_i^{-t_j} \le 1, \tag{4.26}$$

and if $i \neq j$,

$$|\alpha_i - \alpha_j| \ge c(\beta)\max(t_i, t_j), \tag{4.27}$$

where $c(\beta) := 1 - \exp\left(-\frac{1}{2(2\beta+1)^2}\right)$.

Step 3: Computation of the Kullback-Leibler (KL) divergence

In this step, we first compute the KL divergence between F_0 and F_i, which is defined as $KL(F_0, F_i) = \int f_0(x)\log\frac{f_0(x)}{f_i(x)}dx$. Then we prove that it has the same order of the KL divergence between F_i and F_0. Second, we prove that the KL divergence between F_i and F_j is at most of the same order of $\max\{KL(F_0, F_i), KL(F_j, F_0)\}$.

Lemma 5 provides the order of the KL divergence between F_i and F_0.

Lemma 5. *Let $1 \le i \le M$. It holds that for F_0 in (4.23), F_i in (4.24) and using notation in (4.22),*

$$\max\big(KL(F_0, F_i), KL(F_i, F_0)\big) \le \frac{2t_i^2 K_i^{-\alpha}}{\alpha^2}.$$

Using Lemma 5, we obtain bounds on the KL divergence between F_i and F_j in the following lemma.

Lemma 6. *Let $(i, j) \in \{1, \dots, M\}^2$. It holds that for F_i in (4.24) and using notation in (4.22),*

$$KL(F_i, F_j) \le \frac{2\exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2}\left(t_i^2 K_i^{-\alpha} + t_j^2 K_j^{-\alpha}\right). \tag{4.28}$$

Step 4: Use of Fano's method.

Here we follow ideas in the Fano's method using the above results in Step 1-3. Let $\hat{\alpha} = \hat{\alpha}(X_1, \ldots, X_n) =: \hat{\alpha}(X)$ be an estimator of α. Then we define the following discrete random variable

$$Z = Z(X) := \arg \min_{j \in \{1,\ldots,M\}} |\hat{\alpha}(X) - \alpha_j|,$$

which implies that $|\hat{\alpha} - \alpha_j| > c(\beta)t_j/2$ if $Z \neq j$ by Equation (4.27). Also we consider another random variable Y, uniformly distributed on $\{1, \ldots, M\}$ where $X|Y = j \sim F_j^n$. By bounding the maximum by the average,

$$\begin{aligned}
\max_{j \in \{1,\ldots,M\}} \mathbb{P}_{F_j} \left(|\hat{\alpha} - \alpha_j| \geq \frac{c(\beta)t_j}{2} \right) &\geq \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}\left(Z \neq j | Y = j\right) \\
&= \mathbb{P}(Z \neq Y) \\
&\geq 1 - \frac{1}{\log M} \left(\frac{1}{M^2} \sum_{j,j'} KL(F_j^n, F_{j'}^n) + \log 2 \right),
\end{aligned}$$

where the last inequality is obtained by Fano's inequality (see Section 2.1 in Cover and Thomas (2012), or see Appendix for a proof of how this inequality is derived).

Using the fact that $KL(F_1^n, F_2^n) = nKL(F_1, F_2)$, and by Equation (4.28),

$$\begin{aligned}
\frac{1}{M^2} \sum_{j,j'} KL(F_j^n, F_{j'}^n) &\leq \frac{n}{M^2} \frac{2\exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2} \sum_{j,j'} \left(t_j^2 K_j^{-\alpha} + t_{j'}^2 K_{j'}^{-\alpha} \right) = \frac{n}{M} \frac{4\exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2} \sum_j t_j^2 K_j^{-\alpha} \\
&= \frac{n}{M} \frac{4\exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2} \sum_j \frac{\upsilon \log(M)}{n} \\
&= \frac{4\exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2} (\log(M)) \times \upsilon \leq \frac{1}{2} \log(M).
\end{aligned}$$

where the second equality follows by $t_j^2 K_j^{-\alpha} = K_j^{-\alpha(2\beta_j+1)} = \frac{\upsilon \log(M)}{n}$ and the last inequality is by $\upsilon \leq \frac{\alpha^2}{8\exp(\frac{1}{\alpha(2\beta-1)})}$. Hence, for a sufficiently large n, we have

$$\max_{j \in \{1,\ldots,M\}} \mathbb{P}_{F_j} \left(|\hat{\alpha} - \alpha_j| \geq \frac{c(\beta)t_j}{2} \right) \geq \frac{1}{4}.$$

More specifically, using $c(\beta) := 1 - \exp(-\frac{1}{2(2\beta+1)^2}) \geq \frac{1}{2(2\beta+1)^2}$ and since $t_j = \left(\frac{\upsilon \log(M)}{n}\right)^{\frac{\beta_j}{2\beta_j+1}} \geq$

$v^{\frac{\beta_j}{2\beta_j+1}}\Big(\frac{\log\big((\log(n))/2\big)}{n}\Big)^{\frac{\beta_j}{2\beta_j+1}}$, we have

$$\max_{j\in\{1,\ldots,M\}} \mathbb{P}_{F_j}\left(|\hat\alpha-\alpha_j| \geq B(\alpha,\beta,\beta_j)\Big(\frac{\log\big((\log(n))/2\big)}{n}\Big)^{\frac{\beta_j}{2\beta_j+1}}\right) \geq \frac{1}{4},$$

where

$$B(\alpha,\beta,\beta_j) := \frac{1}{4(2\beta+1)^2}\min\left[1, \Big(\frac{\alpha^2}{8\exp(\frac{1}{\alpha(2\beta-1)})}\Big)^{\frac{\beta_j}{2\beta_j+1}}\right]. \tag{4.29}$$

By definition of $\{F_1,\ldots,F_M\}$, we have (by Lemma 4)

$$\{F_1,\ldots,F_M\} \subset \Big\{F\in\mathcal{S}(\alpha^*,\beta^*,C,\tilde C') : \alpha^*\in[\alpha/2,\alpha], \beta^*\in[\beta-1,\beta], C\in[\tilde C_1,\tilde C_2]\Big\},$$

where $\tilde C_1(\alpha,\beta) := \exp\Big(-\frac{1}{\alpha(2\beta-1)}\Big)$, $\tilde C_2 := 1$, and $\tilde C'(\alpha,\beta) = \frac{1}{\alpha(\beta-1)}$.

Then by bounding the supremum by the maximum over the finite subset, we finally provide the following lower bound result.

$$\begin{aligned}\sup_{\substack{\alpha^*\in[\alpha/2,\alpha],\beta^*\in[\beta-1,\beta]\\ C\in[\tilde C_1,\tilde C_2]}}\ \sup_{F\in\mathcal{S}(\alpha^*,\beta^*,C,\tilde C')} &\mathbb{P}_F\left(|\hat\alpha-\alpha^*| \geq B(\alpha,\beta,\beta^*)\Big(\frac{\log\big((\log(n))/2\big)}{n}\Big)^{\frac{\beta^*}{2\beta^*+1}}\right)\\ &\geq \max_{j\in\{1,\ldots,M\}} \mathbb{P}_{F_j}\left(|\hat\alpha-\alpha_j| \geq B(\alpha,\beta,\beta_j)\Big(\frac{\log\big((\log(n))/2\big)}{n}\Big)^{\frac{\beta_j}{2\beta_j+1}}\right)\\ &\geq \frac{1}{4}.\end{aligned}$$

By changing parametrization and setting $\alpha_1 = \alpha/2$ and $\beta_1 = \beta-1$, we proved that

$$\sup_{\substack{\alpha^*\in[\alpha_1,2\alpha_1],\beta^*\in[\beta_1,\infty)\\ C\in[C_1,C_2]}}\ \sup_{F\in\mathcal{S}(\alpha^*,\beta^*,C,C')} \mathbb{P}_F\left(|\hat\alpha-\alpha^*| \geq B_4\Big(\frac{n}{\log\big(\log(n)/2\big)}\Big)^{-\frac{\beta^*}{2\beta^*+1}}\right) \geq 1/4,$$

where $C' = \tilde C'(2\alpha_1,\beta_1+1)$ and

$$C_1 = \tilde C_1(2\alpha_1,\beta_1+1),\ C_2 = 1,\ B_4 = B(2\alpha_1,\beta_1+1,\infty). \tag{4.30}$$

This concludes the proof. □

Proof of Lemma 4. **(1) Proof of Equation** (4.25)**:** For $1 \le i \le M$, $F_i \in \mathcal{A}(\alpha - t_i, K_i^{-t_i})$ by definition. For $x > K_i$, F_i satisfies the second-order Pareto condition. For any $1 \le x \le K_i$,

$$\Big|1 - F_i(x) - K_i^{-t_i} x^{-\alpha+t_i}\Big| = \Big|x^{-\alpha} - K_i^{-t_i} x^{-\alpha+t_i}\Big| = x^{-\alpha}\Big|1 - K_i^{-t_i} x^{t_i}\Big| \\ \le 2x^{-\alpha}\Big|t_i \log(K_i/x)\Big|.$$

The last inequality is obtained since $\forall u \in [0,1]$, $|e^{-u} - 1| \le 2u$ and

$$t_i \log(K_i) \le n^{-\frac{\beta_i}{2\beta_i+1}} \big(\log M\big)^{\gamma_i} \big(\frac{1}{\alpha(2\beta_i+1)}\big) \log(n) \le \frac{1}{\alpha} n^{-\frac{\beta_i}{2\beta_i+1}} \log(n)^{\gamma_i+1} \le 1$$

by assuming large n. Then for any $1 \le x \le K_i$

$$\Big|1 - F_i(x) - K_i^{-t_i} x^{-\alpha+t_i}\Big| \le 2x^{-\alpha} K_i^{-\alpha\beta_i} \log(K_i/x) = 2x^{-\alpha} x^{-\alpha\beta_i} \Big(\frac{K_i}{x}\Big)^{-\alpha\beta_i} \log(K_i/x) \\ \le 2x^{-\alpha} x^{-\alpha\beta_i} \Big(\frac{K_i}{x}\Big)^{-\alpha(\beta-1)} \log(K_i/x) \\ \le \frac{1}{\alpha(\beta-1)} x^{-\alpha(\beta_i+1)},$$

where the ultimate inequality follows from the fact that for any $u \ge 1, t > 0$, we have $u^{-t}\log(u) \le 1/(et)$. Thus, we have shown the first result (4.25).

(2) Proof of Equation (4.26)**:** Let $1 \le j \le M$. Since $K_i > 1$ and $t_j > 0$ for all $i = 1, \dots, M$ and all $j = 1, \dots, M$, we have $K_i^{-t_j} \le 1$. By definition (4.22) and bounding $M \le n$,

$$K_i^{-t_j} \ge \Big(n^{\frac{1}{\alpha(2\beta_i+1)}}\Big)^{-n^{-\frac{\beta_j}{2\beta_j+1}} \big(\log M\big)^{\gamma_j}} = \exp\Big(-\frac{\log(n)}{\alpha(2\beta_i+1)} n^{-\frac{\beta_j}{2\beta_j+1}} \big(\log M\big)^{\gamma_j}\Big) \\ \ge \exp\Big(-\frac{\log(n)^{1+\gamma_j}}{\alpha(2\beta-1)} n^{-\frac{\beta_j}{2\beta_j+1}}\Big) \\ \ge \exp\Big(-\frac{1}{\alpha(2\beta-1)}\Big),$$

where the final inequality follows for a sufficiently large n.

(3) Proof of Equation (4.27)**:** Consider now $i < j$. From (4.25), each F_i corresponds to the tail index $\alpha_i = \alpha - t_i = \alpha - (n/(\upsilon\log(M))^{-\beta_i/(2\beta_i+1)}$. For $i < j$, we have $\alpha_i > \alpha_j$ and $t_i < t_j$ as we described in the Step 1. Also, using $\beta_j - \beta_i = (i-j)/M$,

$$\begin{aligned}
|\alpha_i - \alpha_j| &= \Big|t_j(1 - \frac{t_i}{t_j})\Big| = t_j \Big|1 - (\frac{n}{\upsilon \log(M)})^{-\frac{\beta_i}{2\beta_i+1} + \frac{\beta_j}{2\beta_j+1}}\Big| \\
&= t_j \left[1 - \Big(\frac{n}{\upsilon \log(M)}\Big)^{\frac{(i-j)/M}{(2\beta_i+1)(2\beta_j+1)}}\right] = t_j \left[1 - \exp\Big(\frac{(i-j)}{M(2\beta_i+1)(2\beta_j+1)} \log\big(\frac{n}{\upsilon \log M}\big)\Big)\right] \\
&\geq t_j \left(1 - \exp\Big(\frac{(i-j)}{(2\beta_i+1)(2\beta_j+1)} \frac{(M-1)}{M}\Big)\right) \\
&\geq t_j \left[1 - \exp\left(\frac{(i-j)}{2(2\beta_i+1)(2\beta_j+1)}\right)\right],
\end{aligned}$$

where the penultimate inequality is obtained since $\upsilon \leq 1$, and since $\log\big(\frac{n}{\log(M)}\big) + 1 \geq M \geq 2$. This implies Equation (4.27).

□

Proof of Lemma 5. **(1) KL divergence between F_0 and F_i**

Let $1 \leq i \leq M$. By definition of KL divergence,

$$\begin{aligned}
KL(F_0, F_i) &= \int_1^\infty f_0(x) \log\left(\frac{f_0(x)}{f_i(x)}\right) dx \\
&= -t_i \int_{K_i}^\infty \alpha x^{-\alpha-1} \log\left(\left(\frac{\alpha - t_i}{\alpha}\right)^{\frac{1}{t_i}} \frac{x}{K_i}\right) dx.
\end{aligned}$$

By the change of variable $u = \big(\frac{\alpha - t_i}{\alpha}\big)^{1/t_i} x/K_i$, and letting $a_i = \big(\frac{\alpha - t_i}{\alpha}\big)^{1/t_i}$,

$$\begin{aligned}
KL(F_0, F_i) &= -t_i \int_{a_i}^\infty \alpha\Big((\frac{\alpha}{\alpha - t_i})^{1/t_i} K_i u\Big)^{-\alpha-1} \log(u) du \times \Big((\frac{\alpha}{\alpha - t_i})^{1/t_i} K_i\Big) \\
&= t_i \Big(a_i^{-1} K_i\Big)^{-\alpha} \int_{a_i}^\infty (-\alpha) u^{-\alpha-1} \log(u) du.
\end{aligned}$$

Now by performing an integration by parts, we obtain

$$\begin{aligned}
KL(F_0, F_i) &= t_i \Big(a_i^{-1} K_i\Big)^{-\alpha} \left(u^{-\alpha} \log(u)\big|_{a_i}^\infty - \int_{a_i}^\infty u^{-\alpha-1} du\right) \\
&= t_i K_i^{-\alpha} \Big(\log(1/a_i) - \frac{1}{\alpha}\Big) = K_i^{-\alpha} \left(\log\left(\frac{\alpha}{\alpha - t_i}\right) - \frac{t_i}{\alpha}\right).
\end{aligned}$$

Using $\alpha - t_i \geq \alpha/2$, we further upper bound this divergence

$$\begin{aligned}
KL(F_0, F_i) &= K_i^{-\alpha} \left(\log\left(1 + \frac{t_i}{\alpha - t_i}\right) - \frac{t_i}{\alpha}\right) \leq K_i^{-\alpha} \left(\frac{t_i}{\alpha - t_i} - \frac{t_i}{\alpha}\right) = K_i^{-\alpha} \frac{t_i^2}{\alpha(\alpha - t_i)} \\
&= \frac{2 t_i^2 K_i^{-\alpha}}{\alpha^2}.
\end{aligned}$$

(2) KL divergence between F_i and F_0

Similar calculations as above give

$$\begin{aligned} KL(F_i, F_0) &= \int_1^\infty f_i(x) \log \frac{f_i(x)}{f_0(x)} dx \\ &= t_i a_i^{-\alpha+t_i} K_i^{-\alpha} \int_{a_i}^\infty (\alpha - t_i) u^{-\alpha+t_i-1} \log(u) du \\ &= K_i^{-\alpha} \left(\log\left(\frac{\alpha - t_i}{\alpha}\right) + \frac{t_i}{\alpha - t_i}\right) \le \frac{2t_i^2 K_i^{-\alpha}}{\alpha^2}. \end{aligned}$$

□

Proof of Lemma 6. **(1) KL divergence between F_i and F_j with $i < j$**

Consider the case $i < j$. First, note that

$$\begin{aligned} KL(F_i, F_j) &:= \int f_i(x) \log \frac{f_i(x)}{f_j(x)} dx \\ &= KL(F_i, F_0) + \int_{K_j}^\infty f_i(x) \log \frac{f_0(x)}{f_j(x)} dx. \end{aligned} \tag{4.31}$$

Thus it suffices to bound the second term $\int_{K_j}^\infty f_i \log \frac{f_0}{f_j}$ in (4.31).

We use the similar calculations used in the proof of Lemma 5. With the notation $a_j = (\frac{\alpha - t_j}{\alpha})^{1/t_j}$,

$$\begin{aligned} \int_{K_j}^\infty f_i(x) \log \frac{f_0(x)}{f_j(x)} dx &= t_j K_i^{-t_i} K_j^{-\alpha+t_i} a_j^{\alpha - t_j} \int_{a_j}^\infty -(\alpha - t_i) u^{-\alpha+t_i-1} \log(u) du \\ &= (\frac{K_j}{K_i})^{t_i} K_j^{-\alpha} a_j^{t_i - t_j} \left(\log \frac{1}{a_j} - \frac{1}{\alpha - t_i}\right) \\ &\le 2 \exp\big(\frac{1}{\alpha(2\beta - 1)}\big) t_j^2 K_j^{-\alpha}, \end{aligned}$$

where the final inequality follows by bounding $(K_j/K_i)^{t_i} \le \exp\big(\frac{1}{\alpha(2\beta-1)}\big)$ using Lemma 4, and by bounding $a_j^{t_i - t_j}(\log(1/a_j) - 1/(\alpha - t_i)) \le t_j^2/\alpha^2$ for a sufficiently large n.

Combining this upper bound with bounds on $KL(F_0, F_j)$ and $KL(F_i, F_0)$ in Lemma 5 and also with Equation (4.31),

$$\begin{aligned} KL(F_i, F_j) &\le KL(F_i, F_0) + \exp\left(\frac{1}{\alpha(2\beta - 1)}\right) KL(F_0, F_j) \\ &\le \frac{2 \exp(\frac{1}{\alpha(2\beta-1)})}{\alpha^2} \left(t_i^2 K_i^{-\alpha} + t_j^2 K_j^{-\alpha}\right). \end{aligned} \tag{4.32}$$

(2) KL between F_i and F_j with $i > j$

Now we turn to the case $i > j$. In the same way as for Equation (4.31), we have

$$KL(F_i, F_j) = KL(F_i, F_0) + \int_{K_j}^{\infty} f_i(x) \log \frac{f_0(x)}{f_j(x)} dx. \tag{4.33}$$

For the second term, first note that $\log \frac{f_0(x)}{f_j(x)}$ is a decreasing function for any $x \geq K_j$. Also since $\forall x \geq K_j$, $F_i(x) \leq F_0(x)$, and since $F_i(K_j) = F_0(K_j)$, the measure associated to F_i restricted to $[K_j, \infty)$ stochastically dominates F_0. This implies that

$$\int_{K_j}^{\infty} f_i(x) \log \frac{f_0(x)}{f_j(x)} dx \leq \int_{K_j}^{\infty} f_0(x) \log \frac{f_0(x)}{f_j(x)} dx.$$

Combining this with (4.33) followed by Lemma 5, we have

$$KL(F_i, F_j) \leq KL(F_i, F_0) + KL(F_0, F_j) \leq \frac{2}{\alpha^2} \left(t_i^2 K_i^{-\alpha} + t_j^2 K_j^{-\alpha} \right). \tag{4.34}$$

Finally, by Equations (4.32) and (4.34), we obtain the result (4.28). □

Remark 4. *We only proved the results for certain sets of $C_1, C_2, \beta_1, \beta_2, \alpha_1, \alpha_2, C'$. In fact, it is possible to modify this result to hold for different ranges of parameters (although, the ranges cannot be taken too tight, and C' cannot be taken too small). Note that the narrower the intervals $[C_1, C_2]$, $[\alpha_1, \alpha_2]$, the larger β_1 and the smaller C', the better the result is. Here are possible modification:*

1. *Range of α: from the proof, one could take $[\alpha_1, \alpha_1 + t_M]$ which is actually included in $[\alpha_1, \alpha_1 + n^{-\epsilon}]$ for some $\epsilon > 0$. So without additional effort, the interval can be taken at any position and the range of the interval can be made very small.*
2. *Range of β: for any $\beta_1 > 0$, the result holds for $[\beta_1, \beta_1 + 1]$ (although it is stated for $[\beta_1, \infty)$ to match the upper bound). The constants in the proof could be modified to consider a range $[\beta_1, \beta_1 + \epsilon]$ for any arbitrary small $\epsilon > 0$, by constructing M different β_i's uniformly spread on this interval.*
3. *Range of C: from the proof of the second result in Lemma 4, the tightest range of C is $[K_M^{-t_M}, K_1^{-t_1}]$ which is actually included in $[1 - n^{-\epsilon}, 1]$ for some $\epsilon > 0$. The range could be changed to any $[a - n^{-\epsilon}, a]$ for $a > 0$ by modifying distributions F_i so that the new distrubutions have a domain starting from $a^{-1/\alpha}$ instead of 1 in (4.24). Then, the interval can be taken at any position a and the range of the interval can be made very small.*

However, C' is an upper bound which characterizes the amount of deviation with respect to the Pareto assumption. It cannot be taken too small since if F_i's are too close to F_j's, they can not be distinguished.

5. Appendix

Lemma 7 (Fano's inequality)**.** *Suppose Y is a uniform random variable on $\{1,\ldots,M\}$, and let Z is a random variable of a function of X, where $X|Y=j\sim\mathbb{P}_j$ with $d\mathbb{P}_j/d\nu=p_j$ where ν is the dominating measure. Then*

$$\mathbb{P}\left(Z\neq Y\right)\geq 1-\frac{1}{\log M}\left(\frac{1}{M^2}\sum_{j,j'}KL(\mathbb{P}_j,\mathbb{P}_{j'})+\log 2\right).$$

Proof. Recall the definition of the entropy $H(Y)=-\sum_y p(y)\log p(y)$ for a discrete random variable Y with a probability mass function $p(y)$. Also we denote $H(Y|Z=z)$ by the conditional entropy of Y given $Z=z$, and we define $H(Y|Z)=-\sum_x\sum_y p(y,z)\log p(y|z)$. Following the terminology used in the information theory, we define *information* between Y and Z as the KL divergence between joint distribution and product of the marginal distribution, i.e. $I(Y,Z)=KL(P_{Y,Z},P_Y\times P_Z)$ where we can show that

$$I(Y,Z)=KL(P_{Y,Z},P_Y\times P_Z)=H(Y)-H(Y|Z) \tag{5.1}$$

by splitting the probability distribution. Finally recall that for $Z=Z(X)$, $I(Y,Z)\leq I(Y,X)$.

Consider the event $E=\mathbf{1}\{Z\neq Y\}$. By splitting the probabilities with different order,

$$\begin{aligned}H(E,Y|Z)&=H(Y|Z)+H(E|Y,Z):=(1)\\&=H(E|Z)+H(Y|E,Z):=(2),\end{aligned}$$

where $(1)=H(Y|Z)$ since E becomes a constant given Y and Z. Then we upper bound (2) as follows,

$$\begin{aligned}(2)&=H(E|Z)+H(Y|E,Z)\\&\leq H(E)+H(Y|E,Z)\\&=H(E)+\mathbb{P}(E=0)H(Y|E=0,Y)+\mathbb{P}(E=1)H(Y|E=1,Z)\\&\leq\log 2+\mathbb{P}(Z\neq Y)\log M.\end{aligned}$$

Combining both (1) and (2), we have

$$H(Y|Z)\leq\log 2+\mathbb{P}(Z\neq Y)\log M,$$

in turn,

$$\mathbb{P}(Z\neq Y)\geq\frac{1}{\log M}\left(H(Y|Z)-\log 2\right). \tag{5.2}$$

Now, using the fact (5.1),

$$\begin{aligned} H(Y|Z) &= \log M - I(Y,Z) \\ &\geq \log M - I(Y,X) \\ &= \log M - \int \sum_y p(y)p(x|y) \log \frac{p(y)p(x|y)}{p(x)p(y)} \\ &= \log M - \int \sum_j \frac{1}{M} \mathbf{1}\{y=j\} p(x|y) \log \frac{p(x|y)}{p(x)} \\ &= \log M - \frac{1}{M} \sum_{j=1}^{M} \int p_j(x) \log \frac{p_j(x)}{\frac{1}{M}\sum_{j'} p_{j'}(x)} dx \\ &\geq \log M - \frac{1}{M^2} \sum_{j,j'} KL(\mathbb{P}_j, \mathbb{P}_{j'}), \end{aligned} \tag{5.3}$$

where the penultimate equality is followed since $p(x) = \sum_j \mathbb{P}(Y = j)\mathbb{P}(X = x|Y = j) = \frac{1}{M}\sum_j p_j(x)$, and the last inequality is obtained by the concavity of the logarithm function. Combining (5.2) and (5.3), we obtain

$$\mathbb{P}(Z \neq Y) \geq 1 - \frac{1}{\log M} \left(\frac{1}{M^2} \sum_{j,j'} KL(\mathbb{P}_j, \mathbb{P}_{j'}) + \log 2 \right).$$

□