Jump or Kink: Note on Super-efficiency in Segmented Linear Regression Break-point Estimation

BY YINING CHEN

Department of Statistics, London School of Economics and Political Science
Houghton Street, London WC2A 2AE, U.K.
y.chen101@lse.ac.uk

SUMMARY

We consider the problem of segmented linear regression with a single break-point, with the focus on estimating the location of the break-point. Let \( n \) be the sample size, we show that the global minimax convergence rate for this problem in terms of the mean absolute error is \( O(n^{-1/3}) \). On the other hand, we demonstrate the construction of a super-efficient estimator that achieves the pointwise convergence rate of either \( O(n^{-1}) \) or \( O(n^{-1/2}) \) for every fixed parameter value, depending on whether the structural change is a jump or a kink. The implications of this example and a potential remedy are discussed.

Some key words: change-point; minimax rate; pointwise rate; structural break

1. INTRODUCTION

Asymptotic analysis is commonly used to facilitate comparison between different statistical estimators from a frequentist’s perspective. Once the consistency of an estimator is established, the focus then naturally moves onto its rate of convergence. In general, statements concern the following two types of rates: the pointwise rate where the limit is taken when the unknown parameter is fixed, and the uniform rate where the limit is taken as the supremum over some or all of the parameter space. In addition, the convergence rate of the estimator that achieves the fastest uniform rate among all the estimators is known as the minimax rate. Often the (global) minimax rate is used to characterise the hardness of the problem.

In many settings, the pointwise rate, the uniform rate and the minimax rate are the same, in which case the corresponding estimator is usually regarded as rate-optimal. However, there are exceptions where caution must be exercised. A notable example arises from the phenomenon of super-efficiency, first documented by Joseph L. Hodges, Jr. in 1951. This topic was later treated comprehensively by Le Cam (1953) and Hájek (1972), among many others, in the settings of regular parametric models. See Stigler (2007) and Vovk (2009) for excellent reviews of the turbulent history of early studies. More recently, super-efficiency has also been investigated in other more complicated settings. For instance, Brown et al. (1997) studied it in nonparametric function estimation, Heinrich & Kahn (2018) studied it in mixture models, and Banerjee et al. (2019) studied it in the setting of isotonic regression.

Mathematically, let’s denote the parameter space of interest by \( \Theta \), any estimator of \( \theta \in \Theta \) by \( \hat{\theta} \), the loss function by \( L(\theta, \hat{\theta}) \), and the corresponding risk function by \( R(\theta, \hat{\theta}) = E_\theta L(\theta, \hat{\theta}) \). For every \( \theta \in \Theta \), suppose that there exists some \( \gamma_0 > 0 \) such that

\[
0 < \liminf_{\epsilon \to 0+} \liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta: \|\theta - \theta'\| \leq \epsilon} n^{-\gamma_0} R(\theta', \hat{\theta}) \leq \limsup_{\epsilon \to 0+} \limsup_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta: \|\theta - \theta'\| \leq \epsilon} n^{-\gamma_0} R(\theta', \hat{\theta}) < \infty,
\]

then \( n^{-\gamma_0} \) is known as the local minimax rate at \( \theta \). In this context, an estimator \( \hat{\theta} \) is super-efficient in its convergence rate if

\[
\limsup_{n \to \infty} n^{-\gamma_0} R(\theta, \hat{\theta}) < \infty \quad \text{for every } \theta \in \Theta \quad \text{and} \quad \limsup_{n \to \infty} n^{-\gamma_0} R(\theta, \hat{\theta}) = 0 \quad \text{for some } \theta \in \Theta.
\]
The purpose of this short note is to demonstrate that super-efficiency can occur in the setting of segmented linear regression, even with only a single break-point. In spite of the popularity of segmented regression in statistics and econometrics literatures, to our knowledge, this phenomenon has not been widely understood in these contexts. In particular, since the class of segmented linear regression models is not regular (e.g. not differentiable in quadratic mean), existing results regarding super-efficiency in regular parametric models cannot be immediately applied. By focusing on estimating the location of the single break-point and taking the loss function to be the Euclidean distance between the true location and estimated location of the break-point, we show that the global minimax convergence rate of the risk is at least $O(n^{-1/3})$, i.e.

$$\liminf_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} n^{1/3} R(\theta, \hat{\theta}) > 0.$$ 

We then illustrate super-efficiency by constructing an estimator $\hat{\theta}^S$ that for every fixed $\theta \in \Theta$, depending on whether the break-point is a jump or a kink, satisfies either

$$n R(\theta, \hat{\theta}^S) < \infty \quad \text{or} \quad n^{1/2} R(\theta, \hat{\theta}^S) < \infty.$$ 

These findings point to an interesting and dramatic scenario in which the break-point can be estimated at the rate of $O_p(n^{-1/3})$ in the minimax sense. However, as long as we are willing to assume that the truth does exist (i.e. the location, as a proportion of the data sequence, and the size of change do not vary with $n$), the break-point can then be estimated at a much faster rate of either $O_p(n^{-1})$ or $O_p(n^{-1/2})$. Consequently, for this particular break-point estimation problem, caution must be taken when one uses the global minimax rate to characterise its difficulty, and when one compares and interprets the convergence rates of different estimators.

Here our focus is on the segmented linear regression with a single break-point with the aim of illustrating super-efficiency. These results also hold in the setting of multiple break-points under suitable conditions. There has been a number of recent work on estimating the number and locations of unknown break-points in the settings of both continuous and discontinuous piecewise linear mean signals. See Bai and Perron (1998), Muggeo (2003), Das et al. (2016), Maidstone et al. (2019), Baranowski et al. (2019), to name but a few. In particular, with less stringent spacing conditions between consecutive break-points in the setting of a continuous piecewise linear mean signal, Maidstone et al. (2019) and Baranowski et al. (2019) proposed estimators that could achieve within a logarithmic factor of $O_p(n^{-1/3})$ in estimating the locations of all unknown break-points. The estimator’s convergence rate was further improved to within a logarithmic factor of $O_p(n^{-1/2})$ in Baranowski et al. (2019) under more restrictive assumptions. See also Hansen (2017) for inference in the presence of a kink, Hidalgo et al. (2019) for a test of continuity at the break-point, and Dong (2018) for a related problem on treatment effect evaluation. There have also been work on kink location estimation in various univariate nonparametric regression settings. See Raimondo (1998), Goldenshluger et al. (2006), Cheng and Raimondo (2008), Wishart & Kulik (2010) Wishart (2011) and references therein. Finally, we mention the work of Korostelev & Lepski (2008) who investigated a version of jump location estimation problem with a growing dimension.

2. **Model Setup: Segmented Linear Regression with a Single Break-Point**

Suppose that we observe $(X_{ni}, Y_{ni})$ for $i = 1, \ldots, n$. Consider the fixed design setting where

$$X_{ni} = i/(n+1) \quad Y_{ni} = f_\theta(X_{ni}) + \sigma \epsilon_{ni}$$

for some $\sigma > 0$ and some function $f_\theta : [0, 1] \to \mathbb{R}$. Here $\epsilon_{n1}, \ldots, \epsilon_{nn}$ are independent and identically distributed $N(0, 1)$ random variables. Furthermore, $f_\theta$ is a piecewise linear function indexed by $\theta =$
Under the setup mentioned in Section \ref{sec:setup}, \( (\tau_0, \alpha_0, \alpha_0^+, \alpha_0^-, \beta_0, \beta_0^+) \in \Theta \subset [0, 1] \times \mathbb{R}^4 \) of the form

\[
  f_\theta(x) = \begin{cases} 
    \alpha_0 + \beta_0^-(x - \tau_0) & \text{if } x \in [0, \tau_0], \\
    \alpha_0^+ + \beta_0^+(x - \tau_0) & \text{if } x \in (\tau_0, 1]. 
  \end{cases}
\]

In other words, \( f_\theta \) has a single break-point at \( \tau_0 \), with its linear part over \([0, \tau_0]\) determined by the slope \( \beta_0^- \) and the intercept \( \alpha_0^- \) at \( (\tau_0)^- \), and its linear part over \((\tau_0, 1]\) determined by the slope \( \beta_0^+ \) and the intercept \( \alpha_0^+ \) at \( (\tau_0)^+ \). For simplicity, we have assumed that \( f_\theta \) is left-continuous, so \( f_\theta(\tau_0) = \alpha_0^- \). If \( |\alpha_0^+ - \alpha_0^-| \neq 0 \), then we refer to \( \tau_0 \) as a jump. Otherwise, if \( |\alpha_0^+ - \alpha_0^-| = 0 \) but \( |\beta_0^+ - \beta_0^-| \neq 0 \), then we call \( \tau_0 \) a kink.

To asymptotically analyse the break-point estimator based on \((X_{n1}, Y_1), \ldots, (X_{nn}, Y_n)\), it is common to assume that the actual break-point does not occur too close to the boundary at \( x = 0 \) or \( x = 1 \), and the structural change is “noticeable”, so at least one of the following two quantities, \( |\alpha_0^+ - \alpha_0^-| \) and \( |\beta_0^+ - \beta_0^-| \), is reasonably large. As such, it is natural to restrict ourselves to the parameter space of

\[
  \Theta = \left\{ \theta \in [0, 1] \times \mathbb{R}^4 \middle| \tau_0 \in [\delta, 1 - \delta], \max \left( |\alpha_0^+ - \alpha_0^-|, |\beta_0^+ - \beta_0^-| \right) \geq \delta \right\}
\]

for some fixed but perhaps unknown small \( \delta > 0 \). Here the dependence of \( \Theta \) on \( \delta \) is suppressed.

As mentioned previously, our main focus is on estimating the location of the break-point. In a sense, we treat \( \alpha_0, \alpha_0^-, \beta_0, \beta_0^- \) as nuisance parameters. For any estimator \( \hat{\theta} = (\hat{\tau}_0, \hat{\alpha}_0, \hat{\alpha}_0^-, \hat{\beta}_0, \hat{\beta}_0^-) \), we evaluate its performance based on the estimated break-point’s absolute loss, namely, with the loss function \( L(\theta, \hat{\theta}) = |\tau_0 - \hat{\tau}_0| \) and the risk function \( R(\theta, \hat{\theta}) = E_\theta L(\theta, \hat{\theta}) \). Analogous conclusions could also be made under other losses, such as \( L(\theta, \hat{\theta}) = |\tau_0 - \hat{\tau}_0|^q \) for some \( q > 1 \).

Finally, we remark that this particular fix design is selected with the aim to better connect to the existing change-point detection literature.

3. Minimax Rate of Convergence

First, we investigate the local minimax rate of convergence. We separate the parameter space into two disjoint sets \( \Theta^K \) and \( \Theta \setminus \Theta^K \), where \( \Theta^K \) is the parameter space representing functions with a kink, i.e.

\[
  \Theta^K = \left\{ \theta \in \Theta \middle| \alpha_0^- = \alpha_0^+ \right\}.
\]

**Theorem 1.** Under the setup mentioned in Section 2,

\[
  \lim_{\varepsilon \to 0+} \liminf_{n \to \infty} \inf_{\bar{\theta} \in \Theta} \sup_{\theta' \in \Theta: ||\theta' - \theta|| \leq \varepsilon} n \mathbb{R}(\theta', \hat{\theta}) > 0 \quad \text{for every } \theta \in \Theta \setminus \Theta^K
\]

and

\[
  \lim_{\varepsilon \to 0+} \liminf_{n \to \infty} \inf_{\bar{\theta} \in \Theta} \sup_{\theta' \in \Theta: ||\theta' - \theta|| \leq \varepsilon} n^{1/3} \mathbb{R}(\theta', \hat{\theta}) > 0 \quad \text{for every } \theta \in \Theta^K. \tag{1}
\]

Theorem 1 implies that when \( \tau_0 \) is a jump, the local minimax rate for estimating the location of the break-point in terms of the magnitude of \( \tau_0 - \hat{\tau}_0 \) is at least of \( O_p(n^{-1}) \). However, this rate slows down considerably to \( O_p(n^{-1/3}) \) when \( \tau_0 \) is a kink.

The next corollary concerns the global minimax rate, which immediately follows from Theorem 1.

**Corollary 1.** Under the setup mentioned in Section 2,

\[
  \lim_{n \to \infty} \inf_{\theta' \in \Theta} \sup_{\theta \in \Theta} n^{1/3} \mathbb{R}(\theta, \hat{\theta}) > 0.
\]

Although it is known that when the exact type of the break-point is given a priori, a jump could be estimated at \( O_p(n^{-1}) \) and a kink could be estimated at \( O_p(n^{-1/2}) \), we emphasize that these facts alone are far from implying the minimax rate of \( O_p(n^{-1/3}) \) for \( \tau_0 - \hat{\tau}_0 \) as shown above. It is also interesting to note that the \( O_p(n^{-1/3}) \) rate also appears in Raimondo (1998) who considers a related problem in the
Under the setup mentioned in Section 4, Y. C.
nonparametric setting where there is a jump in the first derivative of a continuous mean. However, the
class of functions he considered in deriving this rate is different from ours.

On the other hand, if we further constrain the parameter space from \( \Theta \) to \( \Theta^K \), then the following
minimax result for kink location estimation in the setting of continuous segmented linear regression holds.

**Theorem 2.** Under the setup mentioned in Section 2,

\[
\lim_{n \to \infty} \inf_{\hat{\theta}} \sup_{\theta \in \Theta^K} n^{1/2} R(\theta, \hat{\theta}) > 0. \tag{2}
\]

Note that \( O_p(n^{-1/2}) \) implied by (2) is faster than \( O_p(n^{-1/3}) \) implied by (1). This seemingly counter-
intuitive difference in the rates is due to the different choices of the parameter spaces in their derivations
and can be explained by examining the proofs of Theorem 1 and Theorem 2 in the online supplementary
materials. Our proofs follow from Le Cam’s two-point method. See Le Cam (1986) or Yu (1997).

To give some intuitions, we first confine ourselves to \( \theta_1 = (1/2, 0, 0, -1, 1) \) and \( \theta_2 = (1/2 + \Delta, -\Delta, \Delta, -1, 1) \) for some small \( \Delta > 0 \), with \( \theta_1, \theta_2 \in \Theta \). Denote the distribution of \( \{Y_1, \ldots, Y_n\} \) using the
data generating process described in Section 2 with \( \theta_1 \) as \( \mathcal{P}_{\theta_1}^n \), and with that \( \theta_2 \) as \( \mathcal{P}_{\theta_2}^n \). Then
break-point estimation could be viewed as the problem of differentiating between \( \mathcal{P}_{\theta_1}^n \) and \( \mathcal{P}_{\theta_2}^n \) based on the
observations, whose hardness is dictated by the squared total variation distance between them. In
the meantime, this squared total variation distance, denoted by \( \|\mathcal{P}_{\theta_1}^n - \mathcal{P}_{\theta_2}^n\|_{TV}^2 \), can be bounded under
suitable conditions as follows, with \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx \) playing a crucial role in characterising
the hardness of the original problem:

\[
\|\mathcal{P}_{\theta_1}^n - \mathcal{P}_{\theta_2}^n\|_{TV}^2 \leq 2 - 2 \prod_{i=1}^n \left[ 1 - d_{hel}^2 \left\{ N \left( \frac{f_{\theta_1}(i/(n+1)), \sigma^2}{}, N \left( \frac{f_{\theta_2}(i/(n+1)), \sigma^2}{}, \right) \right) \right\} \right]
\]

\[
= 2 - 2 \exp \left[ -\frac{1}{8\sigma^2} \sum_{i=1}^n \left\{ f_{\theta_1}(i/(n+1)) - f_{\theta_2}(i/(n+1)) \right\}^2 \right]
\]

\[
\rightarrow 2 - 2 \exp \left[ -\frac{n}{8\sigma^2} \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx \right].
\]

Here \( d_{hel} \) is the Hellinger distance and \( d_{hel}^2 \left\{ N(\mu_1, \sigma), N(\mu_2, \sigma) \right\} = 1 - \exp \left\{ -(\mu_1 - \mu_2)^2/(8\sigma^2)^2 \right\}. \) For
this particular pair of \( \theta_1 \) and \( \theta_2 \), \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx = O(\Delta^2) \) for \( \Delta \to 0 \). See Fig. 1(a). Here
\( f_{\theta_1}(x) \) and \( f_{\theta_2}(x) \) only differ over \( x \in (1/2, 1/2 + \Delta) \), meaning that the problem can be viewed as a
local one as most pairs of the observations, namely, \( (X_n, Y_n) \) with \( X_n \notin (1/2, 1/2 + \Delta) \), are irrelevant.
In contrast, with the same value of \( \theta_1 \), if we only consider parameters in \( \Theta^K \) we are then unable to
find a \( \theta_2 \in \Theta^K \) with the corresponding break-point at \( 1/2 + \Delta \) such that \( f_{\theta_1}(x) \) and \( f_{\theta_2}(x) \) only differ
over a small neighbourhood. In fact, \( f_{\theta_1}(x) \) and \( f_{\theta_2}(x) \) will have to differ over a substantial interval that
does not shrink as \( \Delta \to 0 \), so the problem of distinguishing between \( f_{\theta_1} \) and \( f_{\theta_2} \) appears more global
than before. By taking, for example, \( \theta_2 = (1/2 + \Delta, -\Delta, -\Delta, -1, 1) \) (while keeping the same \( \theta_1 \)), we
obtain \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx = O(\Delta^2) \), as demonstrated in Fig. 1(b). In fact, we can further show that
\( \inf_{\theta_2 \in \Theta^K} \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx = O(\Delta^2) \). This order difference of \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx \)
in \( \Delta \) implies that the cases of \( \theta \in \Theta \) and \( \theta \in \Theta^K \) are fundamentally different! To give more details, in Le
Cam’s method, the minimax rate can be derived by picking \( \Delta \) such that \( \|\mathcal{P}_{\theta_1}^n - \mathcal{P}_{\theta_2}^n\|_{TV} \leq C \) for some
constant \( C < 1 \). In our settings, as derived as above, this roughly amounts to requiring \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx = O(n^{-1}) \). As such, \( \Delta \) would be taken as \( O(n^{-1/3}) \) in Fig. 1(a), and \( O(n^{-1/2}) \) in Fig. 1(b),
which are also the minimax convergence rates for break-point estimation under \( \theta \in \Theta \) and \( \theta \in \Theta^K \) re-
spectively in terms of the expected absolute loss. Finally, for completeness, we also illustrate the case
of a “noticeable” jump in Fig. 1(c), where \( \theta_1 = (1/2, -1/2, 1/2, -1, 1) \) (i.e. with jump size of 1) and
\( \theta_2 = (1/2 + \Delta, -1/2 - \Delta, 1/2 + \Delta, -1, 1) \). Since here \( \int_0^1 \{f_{\theta_1}(x) - f_{\theta_2}(x)\}^2 dx = O(\Delta) \), the mini-
max rate for estimating break-point of this type is \( O(n^{-1}) \).
4. A SUPER-EFFICIENT ESTIMATOR

Write $\Theta = [0,1] \times \mathbb{R}^4$ and $\Theta^K = \{ \theta \in \Theta | \alpha^- = \alpha^+ \}$. For notational convenience, in Section 4 and Section 5, $\hat{\Theta}$ is denoted as the least squares estimator satisfying

$$\hat{\Theta} := \hat{\Theta}_{LS} \in \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \left( Y_{ni} - f_{\theta}(X_{ni}) \right)^2.$$

Here we minimise over $\Theta$ instead of $\Theta$, because $\delta$ of $\Theta$ is not always known a priori. Equivalently, we can write

$$\tau_{\delta} \in \arg\min_{\tau \in [0,1]} \left\{ \min_{\alpha, \beta} \sum_{i=1}^{n} (Y_{ni} - \beta X_{ni} - \alpha)^2 1_{\{X_{ni} \in [0, \tau]\}} + \min_{\alpha, \beta} \sum_{i=1}^{n} (Y_{ni} - \beta X_{ni} - \alpha)^2 1_{\{X_{ni} \in (\tau, 1]\}} \right\}.$$

Similarly, for kink estimation where we further restrict ourselves to $\Theta^K$, we denote the corresponding least squares estimator by $\hat{\Theta}^K \in \arg\min_{\theta \in \Theta^K} \sum_{i=1}^{n} \left( Y_{ni} - f_{\theta}(X_{ni}) \right)^2$, with

$$\tau_{\delta} \in \arg\min_{\tau \in [0,1]} \left[ \min_{\alpha, \beta^-, \beta^+} \sum_{i=1}^{n} \left( Y_{ni} - \alpha - \beta^- (X_{ni} - \tau) 1_{\{X_{ni} \in [0, \tau]\}} - \beta^+ (X_{ni} - \tau) 1_{\{X_{ni} \in (\tau, 1]\}} \right)^2 \right].$$
For uniqueness, we take \( \tau_0 \) and \( \tau_{0}^{\Theta} \) to be the smallest element in the respective sets of minima.

As we shall see, \( \tau_{0}^{\Theta} \) achieves \( O_p(n^{-1}) \) for fixed \( \theta \in \Theta \setminus \Theta^K \), but slows down to \( O_p(n^{-1/3}) \) when \( \theta \in \Theta^K \). Meanwhile, \( \tau_{0}^{\Theta} \) achieves \( O_p(n^{-1/2}) \) for \( \theta \in \Theta^K \). Therefore, making the correct extra assumption of continuity at the break-point and using the corresponding estimator could improve the convergence rate from \( O_p(n^{-1/3}) \) to \( O_p(n^{-1/2}) \) for \( \theta \in \Theta^K \). This motivates us to shrink \( \theta \) towards \( \Theta^K \) in certain cases to improve the pointwise rate. In particular, we could estimate the break-point by \( \tau_{0}^{\Theta^{b}_S} \), where

\[
\hat{\theta}^S = \begin{cases} 
\hat{\theta}^K & \text{if } |\alpha^+_{\hat{\theta}} - \alpha^-_{\hat{\theta}}| \leq n^{-1/6} \\
\hat{\theta} & \text{if } |\alpha^+_{\hat{\theta}} - \alpha^-_{\hat{\theta}}| > n^{-1/6}.
\end{cases}
\]

(4)

We are now in the position to discuss the pointwise and local uniform convergence rates of \( \tau_{0}^{\Theta^{b}_S} \).

**Theorem 3.** Under the setup mentioned in Section 2, \( \tau_{0}^{\Theta^{b}_S} \) is a super-efficient estimator for \( \tau_0 \). In particular, we have

\[
\lim_{n \to \infty} \sup_n nR(\theta, \hat{\theta}^S) < \infty \quad \text{for every } \theta \in \Theta \setminus \Theta^K
\]

and

\[
\lim_{n \to \infty} n^{1/2}R(\theta, \hat{\theta}^S) < \infty \quad \text{for every } \theta \in \Theta^K.
\]

It follows from Theorem 3 that \( \sup_{\theta \in \Theta} \lim_{n \to \infty} n^{1/2}R(\theta, \hat{\theta}^S) < \infty \), i.e. the pointwise rate of break-point estimation via \( \hat{\theta}^S \) for every \( \theta \in \Theta \) is faster than the global minimax rate of \( O(n^{-1/3}) \).

**Theorem 4.** Under the setup mentioned in Section 2, we have that

\[
\lim_{\epsilon \to 0^+} \limsup_{n \to \infty} \sup_{\theta' \in \Theta \setminus \Theta^K, \|\theta' - \theta\| \leq \epsilon} nR(\theta', \hat{\theta}^S) < \infty \quad \text{for every } \theta \in \Theta \setminus \Theta^K
\]

and

\[
\lim_{\epsilon \to 0^+} \liminf_{n \to \infty} \sup_{\theta' \in \Theta \setminus \Theta^K, \|\theta' - \theta\| \leq \epsilon} n^{1/3}R(\theta', \hat{\theta}^S) = \infty \quad \text{for every } \theta \in \Theta^K.
\]

Theorem 4 implies that the global uniform rate of \( \tau_{0}^{\Theta^{b}_S} \) in terms of the absolute loss is worse than the global minimax rate of \( O(n^{-1/3}) \), which could actually be achieved by \( \tau_{0}^{\Theta^{b}_S} \). This type of behaviour is typical for super-efficient estimators that tend to achieve better pointwise convergence rates at the cost of worse uniform convergence rates.

In addition, the construction of the super-efficient estimator is by no means unique. In fact, here the threshold \( n^{-1/6} \) in (4) can be replaced by \( cn^{-\gamma} \) for any fixed \( c > 0 \) and \( \gamma \in (0, 1/3] \). Alternatively, one could replace \( |\alpha^+_{\theta} - \alpha^-_{\theta}| \) in (4) by the difference between residual sum of squares from fitting the model over either \( \Theta^K \) or \( \Theta \), and then choose the cut-off decision boundary accordingly.

5. Numerical Experiment

We run a small simulation study to compare the behaviour of \( \tau_0 \) and \( \tau_{0}^{\Theta^{b}_S} \). Two different scenarios are considered under the settings of Section 2:

(a) \( \theta_1 = (0.5, 0, 0, -1, 1) \in \Theta^K \), i.e. \( f_{\theta_1}(x) = |x - 0.5| \);

(b) \( \theta_2 = (0.5, 0, 0.5, -1, 1) \in \Theta \setminus \Theta^K \), i.e. \( f_{\theta_2}(x) = |x - 0.5| + 1_{\{x > 0.5\}} \).

Here we take \( \sigma = 0.5 \) and \( n = 100, 200, 500, 1000, 2000 \). All experiments are repeated 1000 times. The estimated values of \( R(\theta, \hat{\theta}) \) and \( R(\theta, \hat{\theta}^S) \), also known as the mean absolute errors (MAEs) of \( \tau_0 \) and \( \tau_{0}^{\Theta^{b}_S} \), are reported in Fig. 2 on a log-log scale.

In Fig. 2(a), the super-efficiency phenomenon is visible, where the super-efficient estimator \( \tau_{0}^{\Theta^{b}_S} \) performs better than the least squares estimator \( \tau_0 \) in the presence of a kink, especially for large \( n \). It is also
evident from the plot that $\tau_{\theta^b}$ and $\tau_{\theta^s}$ have different pointwise convergence rates there, as indicated in Section 4. Meanwhile, Fig. 2(b) demonstrates that in the presence of a jump, $\tau_{\theta^b}$ and $\tau_{\theta^s}$ perform similarly. In particular, for large $n = 2000$, they are exactly the same in all the 1000 runs. Finally, Fig. 2 confirms that in terms of pointwise rates, estimating the location of a jump is easier than estimating that of a kink.

---

**Fig. 2.** Estimated MAEs of $\tau_{\theta^b}$ and $\tau_{\theta^s}$, for $n = 100, 200, 500, 1000, 2000$ on a log-log scale under different scenarios: (a) $\theta_1$ with a kink; (b) $\theta_2$ with a jump.

---

6. DISCUSSION

Although the super-efficient estimator $\tau_{\theta^s}$ achieves a pointwise rate faster than the global minimax rate for every $\theta \in \Theta$, it is clear that in our example super-efficiency only occurs over $\Theta^K$, which is a Lebesgue null set in comparison to $\Theta$. However, if we were to focus solely on $\tau_{\theta^b}$ by treating $\alpha_\theta^+, \alpha_\theta^-, \beta_\theta^+$ and $\beta_\theta^-$ as nuisance parameters, then super-efficiency could occur at every $\tau_\theta \in [\delta, 1 - \delta]$. Here $[\delta, 1 - \delta]$ is no longer a Lebesgue null set in comparison to $[0, 1]$.

On the other hand, whilst $\tau_{\theta^b}$ achieves a better pointwise convergence rate for locating the break-point, like Hodges' estimator, it is penalised at local neighbouring points. In particular, $\limsup_{n \to \infty} \sup_{\theta \in \Theta} n^{1/3} R(\theta, \hat{\theta}^b) = \infty$, but $\limsup_{n \to \infty} \sup_{\theta \in \Theta} n^{1/3} R(\theta, \hat{\theta}) < \infty$, where $\hat{\theta}$ is taken as the least squares estimator defined in (3). Thus, in terms of the uniform convergence rate, $\tau_{\theta^s}$ actually performs worse than $\tau_{\theta^b}$.

However, it is useful to think about whether this perspective of uniformity is what we are really interested in. In this break-point estimation problem, the global minimax convergence rate is derived by considering alternatives with the jump size $n^{1/2}$, but $\limsup_{n \to \infty} \sup_{\theta \in \Theta} n^{1/3} R(\theta, \hat{\theta}) < \infty$, where $\hat{\theta}$ is taken as the least squares estimator defined in (3). Thus, in terms of the uniform convergence rate, $\tau_{\theta^s}$ actually performs worse than $\tau_{\theta^b}$.

It is entirely possible that these alternatives might violate the modeller’s real brief in practice, whose intention dictates that whenever there is a jump or a change in slope (or both), the corresponding change size has to be significant. Mathematically, this plausibly more appropriate parameter space would be a restricted version of $\Theta$, given by

$$\Theta^* = \left\{ \theta \in \Theta \mid \min \left( \alpha_\theta^+ - \alpha_\theta^- \right) \geq \delta \right\}.$$ 

It can then be shown that for every $\theta \in \Theta^K$,

$$\liminf_{\epsilon \to 0^+} \liminf_{n \to \infty} \inf_{\theta \in \Theta^* : \|\theta - \theta^s\| \leq \epsilon} \sup n^{1/2} R(\theta^s, \hat{\theta}) > 0,$$

implying that the local minimax convergence rate for estimating the kink over $\Theta^*$ is $O_p(n^{-1/2})$, and thus $\hat{\theta}^s$ is no longer super-efficient over $\Theta^*$.

Besides, one interesting feature of this break-point estimation problem is that the local minimax convergence rates are different across the parameter space. In this circumstance, it might not be entirely adequate to summarise the hardness of the problem by a single global minimax rate. See also Donoho et al. (1995).
Our example can be generalised in more complex settings, such as segmented polynomial regression with higher order polynomials, as well as those with heterogeneous sub-Gaussian errors and multiple break-points. It shows the importance of choosing suitable parameter spaces for the calculation of uniform or minimax convergence rates. We hope that it also demonstrates the need of interpreting different types of convergence rates and the practical meaning of rate-optimality with care and caution.

**ACKNOWLEDGEMENT**

We thank Moulinath Banerjee, Haeran Cho and Piotr Fryzlewicz for their helpful comments on an earlier draft. We also thank the editors and three anonymous reviewers for their valuable suggestions.

**SUPPLEMENTARY MATERIAL**

Supplementary material available at Biometrika online contains proofs of all the theoretical results.

**REFERENCES**


