Electronic Journal of Statistics
Vol. 16 (2022) 1153–1252
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1927

On robust learning in the canonical change point problem under heavy tailed errors in finite and growing dimensions

Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov

Department of Statistics, University of Michigan
Ann Arbor, Michigan, USA
e-mail: mdeb@umich.edu

Abstract: This paper presents a number of new findings about the canonical change point estimation problem. The first part studies the estimation of a change point on the real line in a simple stump model using the robust Huber estimating function which interpolates between the ℓ_1 (absolute deviation) and ℓ_2 (least squares) based criteria. While the ℓ_2 criterion has been studied extensively, its robust counterparts and in particular, the ℓ_1 minimization problem have not. We derive the limit distribution of the estimated change point under the Huber estimating function and compare it to that under the ℓ_2 criterion. Theoretical and empirical studies indicate that it is more profitable to use the Huber estimating function (and in particular, the ℓ_1 criterion) under heavy tailed errors as it leads to smaller asymptotic confidence intervals at the usual levels compared to the ℓ_2 criterion. We also compare the ℓ_1 and ℓ_2 approaches in a parallel setting, where one has m independent single change point problems and the goal is to control the maximal deviation of the estimated change points from the true values, and establish rigorously that the ℓ_1 estimation criterion provides a superior rate of convergence to the ℓ_2, and that this relative advantage is driven by the heaviness of the tail of the error distribution. Finally, we derive minimax optimal rates for the change plane estimation problem in growing dimensions and demonstrate that Huber estimation attains the optimal rate while the ℓ_2 scheme produces a rate sub-optimal estimator for heavy tailed errors. In the process of deriving our results, we establish a number of properties about the minimizers of compound Binomial and compound Poisson processes which are of independent interest.

Keywords and phrases: Change point estimation, heavy tailed error, robust learning.

Received June 2021.

Contents

1. Introduction

In the canonical change-point or change-boundary estimation problem, one posits a regression (or a classification) model in which the conditional distribution of the response given the covariate(s) changes from a constant value on one side of an unknown boundary in covariate space to another on the opposite side. Within the genre of regime change problems, the canonical model is a particularly convenient formulation for investigating the fundamentals of estimation and inference, and the challenges involved therein. In particular, in the one-dimensional case, this gives us the so-called 'stump model':

$$Y = \alpha_0 \mathbb{1}_{X \leq d_0} + \beta_0 \mathbb{1}_{X > d_0} + \xi$$

with $\alpha_0 \neq \beta_0$, where X assumes values in $\mathbb{R}$. In the multidimensional scenario with a p-dimensional covariate X, a natural extension is given by

$$Y = \alpha_0 \mathbb{1}_{\psi(X,d_0)\leq 0} + \beta_0 \mathbb{1}_{\psi(X,d_0)>0} + \xi\,,$$

where $\psi(X, d_0) = 0$ defines a low dimensional smooth surface in $\mathbb{R}^p$.

This paper deals with the estimation of change parameters in such models under different estimating functions in both fixed and growing dimensions along with the calibration of minimax optimal rates. The use of a variety of robust estimating functions is necessitated by the fact that heavy-tailed errors frequently drive data generating mechanisms associated with change-point problems in applications pertaining to finance ([10]), hydrology ([5]), climate and environmental science ([41]), internet data ([18]) and genetics ([37]). We show in this paper that such robust criteria are essential for attaining optimal convergence rates when the number of parameters diverges with sample size. We also show that in the fixed dimension scenario the choice of the criterion function does not affect the convergence rate but *does affect* the tails of the limit distribution of the estimated change-point in a way that makes the use of robust criteria more profitable for thick-tailed errors.

We next focus on the organization of the manuscript and articulate the contributions of each section. But before that, we take a moment to introduce the (scaled) Huber estimating function (HEF) ([21]) which is referred to below and used throughout the manuscript. The scaled HEF is defined as $\tilde{H}_k(x) := ((k+1)/k)H_k(x)$ where:

$$H_k(x) = \begin{cases} \frac{x^2}{2}\,. & \text{if } |x| \leq k \\ k\left(|x| - \frac{k}{2}\right)\,, & \text{otherwise}\ . \end{cases}$$

The cost function corresponding to $\tilde{H}_k$ in a generic statistical problem can be written as $C_k(Z,\theta) := \tilde{H}_k(g(Z) - h(Z,\theta))$ where $g(Z)$ is some functional of the data vector Z (say, the real-valued response in a regression model) and $h(Z,\theta)$ is some known function of Z and the parameter θ (say, the regression function). Note that as $k \to 0$, we have $\tilde{H}_k(x) \to |x|$ and for $k \to \infty$, $\tilde{H}_k(x) = x^2/2$, therefore C_k interpolates between the ℓ_1 and ℓ_2 cost functions via the parameter k. The function H_k was introduced in the pioneering work of Peter Huber [20] for the robust estimation of parameters in presence of outliers. The key idea here is the observation that ℓ_1 discrepancy is more robust to outliers than the ℓ_2 discrepancy, whereas ℓ_2 discrepancy has other attractive features like differentiability with constant curvature. The Huber function seeks to combine these two discrepancies and utilize the best of both worlds.

Section 2 presents a treatment of the canonical stump model with a one dimensional covariate under HEF optimization as well as its limiting incarnations (the ℓ_1 and the ℓ_2 criteria) and provides explicit statements of asymptotic distributions which are seen to be the minimizers of various compound Poisson processes. While the limiting behavior under ℓ_2 has been long known in the literature, the study of the asymptotic properties under robust criteria is new.

More interestingly, we are able to characterize the tail behaviors of the limit distributions in terms of the tail-indices of the corresponding error distributions which, to the best of our knowledge, was previously unknown. We demonstrate that under the ℓ_2 criterion the tail index of the error adversely affects the tail of the minimizer of the corresponding compound Poisson process: errors with polynomial decay of tails lead to polynomially decaying tails for the limit; while under HEF (including the ℓ_1 criterion) the tail of the limit distribution is *unaffected* by the tail of the error and is necessarily sub-exponential. This has direct implications for the construction of asymptotic confidence intervals as we discuss later.

Section 3 explores the canonical problem for a growing number of change-point parameters. The first part pertains to situations where multiple change-point parameters are estimated in parallel from *separate* data-sources, and this number is allowed to grow with the total sample size. The second version is the change-boundary problem alluded to at the beginning of our narrative. We explore, specifically, the case of a linear boundary, i.e. a model of the form

$$Y = \alpha_0 \mathbb{1}_{X^T d_0 \leq 0} + \beta_0 \mathbb{1}_{X^T d_0 > 0} + \xi \,,$$

with $\|d_0\|_2 = 1$ (to enforce identifiability) and a p-dimensional covariate X. This is the so-called change-plane model which captures the core features of the change-boundary problem. Our motivation for studying change plane problems stems from the recent use of change-plane models in personalized medicine and related problems [40], [15], as well as the use of change-plane models in econometrics (e.g. see [36], [27], [32] and references therein). We assume that n i.i.d. observations are available from this model and that either $p = o(n)$ or $p \gg n$ with the number of non-zero co-ordinates of d_0 constrained to be appropriately small. We show that in both the parallel change point and high dimensional change plane problems, the ℓ_2 criterion based estimator suffers from the curse of dimensionality unlike its robust counterparts.

Section 4 presents a range of simulation studies in the 1-dimensional case that compare the quantiles of the limit distributions under ℓ_1 and ℓ_2 criteria and discusses the observed patterns. Section 5 concludes, providing among other things an exposition of future challenges in this area.

2. Robust change point estimation in one dimension

Summary: *We analyze the canonical change point model (equation (2.1)) in one dimension under the Huber estimating function $\tilde{H}_k$. The asymptotic distribution of the estimators are presented in Theorem 2.1 - Theorem 2.3. Furthermore, in Theorem 2.4 we show that the tail of the limiting distribution of the least squares estimator is affected by the tail of the error distribution, i.e. a heavy tailed error distribution translates to a heavy tailed limiting distribution, whereas for the least absolute deviation estimator, the limiting distribution has sub-exponential tail irrespective of the tail of the error distribution.*

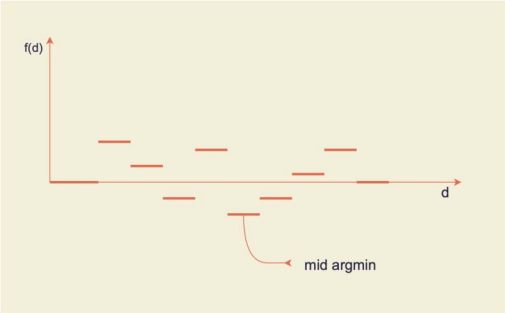

FIG 1. *The mid-argmin of a piecewise constant function*

In this section we analyze the following canonical change point model in one dimension:

$$Y_i = \alpha_0 \mathbb{1}_{X_i \le d_0} + \beta_0 \mathbb{1}_{X_i > d_0} + \xi\,, \tag{2.1}$$

for $1 \le i \le n$. The least squares estimators of the parameters of this model have been well-explored in the literature, but quite surprisingly, nothing is known about its robust variant, and the trade-offs between the two approaches. To understand the difference, consider an even simpler model:

$$Y_i = \mathbb{1}_{X_i > d_0} + \xi_i\,.$$

where $X_i \in \mathbb{R}$ is a real covariate and ξ_i is a mean 0 error independent of X_i. Here d_0 is the parameter of interest, i.e. the *change point* in the space of covariates. Traditionally, one minimizes the squared-error loss to obtain an estimator of d_0:

$$\begin{aligned}
\hat{d}^{\ell_2} &= \text{mid argmin}_{d \in I} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbb{1}_{X_i > d}\right)^2 \\
&= \text{mid argmin}_{d \in I} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \frac{1}{2}\right) \mathbb{1}_{X_i \le d} \\
&:= \text{mid argmin}_{d \in I}\ f(d)\,.
\end{aligned}$$

for some compact interval $I \subset \mathbb{R}$. Note that the function $f(d)$ is a right continuous step function with respect to d, therefore its minimizer is not unique, in fact it is an interval. By mid argmin, we denote the midpoint of the corresponding interval. (See Figure 1 for an illustration.) The statistical properties of this estimator are well-known; e.g. see Chapter 14 of [25] or Proposition 1 of [26] and its preceding discussion. For example, if X is compactly supported with density bounded away from 0 and ∞ on its support, then:

$$n\left(\hat{d}^{\ell_2} - d_0\right) \overset{\mathscr{L}}{\Longrightarrow} \text{mid argmin}_{t \in \mathbb{R}} M(t)$$

where $M(t)$ is a two-sided compound Poisson process with drift described thus: Let $N(t)$ be a homogeneous Poisson process with intensity parameter $f_X(d_0)$

on $[0,\infty)$ where $f_X(\cdot)$ is the density of X. Define two independent stochastic processes $V^+(t)$ on $[0,\infty)$ and $V^-(t)$ on $(-\infty,0]$ as follows:

$$V^+(t) = \sum_{i=1}^{N_1(t)} \left(\xi_i + \frac{1}{2}\right)$$
$$V^-(t) = \sum_{i=1}^{N_2(-t)} \left(\xi_{-i} - \frac{1}{2}\right)$$

where $\{\xi_i\}_{i\in\mathbb{Z}\setminus\{0\}}$ are i.i.d. from the distribution of ξ and $N_1(t), N_2(t)$ are i.i.d copies of $N(t)$, and are independent of the ξ_i's. Then

$$M(t) = V^+(t)\mathbb{1}_{t\geq 0} - V^-(t)\mathbb{1}_{t<0}\,,$$

is a two sided compound Poisson process on the real line (we denote it by $CPP(\xi+1/2, f_X(d_0))$ that drifts off to ∞ on either side, and is minimized almost surely on an interval of points. Taking the mid-argmin of this process ensures symmetry of the limiting distribution under the symmetry of the distribution of ξ.

The asymptotics above require only a second moment for the errors and therefore are valid for many heavy-tailed errors. However, heavy tailed errors enlarge the spread of the limit distribution, resulting in wider confidence intervals for the change-point parameter. This is because the compound Poisson process is closely related to the two sided random walk on $\mathbb{Z}$ with step distribution given by $(\xi + 1/2)$ to the right of 0 and $(-\xi + 1/2)$ to its left. We quantify later in this section how the tail of the distribution of the minimizer of this compound Poisson process depends on the tail index of the error distribution with heavy tailed errors corresponding to a heavier tail for the minimizer which, in turn, implies a wider asymptotic confidence interval.

The natural question, then, is what happens if one were to compute d_0 via the robust HEF, in particular, say the ℓ_1 criterion, and whether asymptotic efficiency relative to the ℓ_2 criterion would accrue as a result in the case of heavy-tailed errors. So, consider:

$$\hat{d}^{\ell_1} = \text{mid argmin}_{d\in I} \frac{1}{n}\sum_{i=1}^{n} |Y_i - \mathbb{1}_{X_i\geq d}|\ .$$

For consistency of $\hat{d}^{\ell_1}$ we need the assumption that $\mathsf{med}(\xi) = 0$. Since the ℓ_2 criterion requires $\mathbb{E}(\xi) = 0$, *the rest of the paper will be developed for symmetric errors* which simplifies the discussion without compromising conceptual issues. We show later (see Theorem 2.2) that

$$n\left(\hat{d}^{\ell_1} - d_0\right) \overset{\mathscr{L}}{\Longrightarrow} \text{mid argmin}_{t\in\mathbb{R}} M_R(t)$$

where $M_R(t)$ is, again, a two sided compound Poisson process with intensity parameter $f_X(d_0)$ and the step-distribution given by that of $|\epsilon+1| - |\epsilon|$. Observe that the random variable $|\epsilon+1| - |\epsilon|$ is bounded in absolute value by 1

irrespective of the tail index of the error and consequently sub-gaussian. This translates to a sub-exponential tail for the asymptotic distribution, resulting in a tighter asymptotic confidence interval than the one obtained via minimizing squared error loss.

We next present our main results for more general stump model described in equation (2.1). Minimizing the Huber estimating function yields the following estimator:

$$\left(\hat{\alpha}^k, \hat{\beta}^k, \hat{d}^k\right) = \text{mid argmin}_{\alpha,\beta,d} \frac{1}{n} \sum_{i=1}^{n} \tilde{H}_k \left(Y_i - \alpha \mathbb{1}_{X_i \leq d} - \beta \mathbb{1}_{X_i > d}\right) \tag{2.2}$$

where, as mentioned earlier, we consider the midpoint of the minimizing interval of d_0. We next present the asymptotic distributions of $(\hat{\alpha}^k, \hat{\beta}^k, \hat{d}^k)$ upon proper centering and scaling.

Theorem 2.1. *Suppose $\theta_0 = (\alpha_0, \beta_0, d_0) \in I$ for some compact subset $I \subset \mathbb{R}^3$. Assume that the density of X is continuous and strictly positive at d_0. Then the estimators $(\hat{\alpha}^k, \hat{\beta}^k, \hat{d}^k)$ obtained in equation (2.2) are asymptotically independent and satisy:*

$$\begin{aligned}
\sqrt{n}(\hat{\alpha}^k - \alpha_0) &\overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{\sigma_k^2}{\mu_k^2 F_X(d_0)}\right), \\
\sqrt{n}(\hat{\beta}^k - \beta_0) &\overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{\sigma_k^2}{\mu_k^2 \bar{F}_X(d_0)}\right), \\
n(\hat{d}^k - d_0) &\overset{\mathscr{L}}{\Longrightarrow} \textit{mid argmin}_{t \in \mathbb{R}} CPP\left(\tilde{H}_k\left(\xi + |\alpha_0 - \beta_0|\right) - \tilde{H}_k(\xi), f_X(d_0)\right)
\end{aligned}$$

where the parameters μ_k and σ_k are:

$$\begin{aligned}
\mu_k &= \frac{k+1}{k} \mathbb{P}\left(-k \leq \xi \leq k\right) \\
\sigma_k^2 &= \left(\frac{k+1}{k}\right)^2 \left(\mathbb{E}\left[\xi^2 \mathbb{1}_{-k \leq \xi \leq k}\right] + 2k^2 \mathbb{P}\left(\xi > k\right)\right).
\end{aligned}$$

where F_X is the distribution of X and $\bar{F}_X$ is $1 - F_X$ is the tail of the distribution and f_X is the density of X.

Note that if $k \to 0$, then $\mu_k \to \mu^{\ell_1} = 2f_\xi(0)$ and $\sigma_k^2 \to (\sigma^{\ell_1})^2 = 1$. One the other hand, if $k \to \infty$, then $\mu_k \to \mu^{\ell_2} = 1$ and $\sigma_k^2 \to (\sigma^{\ell_2})^2 = \sigma_\xi^2$, as long as $E(\xi^2)$ is finite, which is a requirement for the ℓ_2 based estimation strategy to work. The following two theorems present the asymptotic distribution of the estimated parameters (upon proper centering and scaling) for these special cases: ℓ_1 and ℓ_2 criteria, where we see that the limiting parameters are indeed $\mu^{\ell_1}, \sigma^{\ell_1}$ and $\mu^{\ell_2}, \sigma^{\ell_2}$ respectively. We note that the proofs do not directly follow by taking the limit of k in the proof of Theorem 2.1, but rely on similar techniques.

Theorem 2.2. *Consider minimizing the ℓ_1 criterion function to obtain:*

$$\hat{\theta}^{\ell_1} = \underset{\theta \in I}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left|Y_i - \alpha \mathbb{1}_{X_i \leq d} - \beta \mathbb{1}_{X_i > d}\right|$$

Then, under the assumptions of Theorem 2.1, we have:

$$\sqrt{n}(\hat{\alpha}^{\ell_1} - \alpha_0) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{1}{4f_\xi^2(0)F_X(d_0)}\right),$$
$$\sqrt{n}(\hat{\beta}^{\ell_1} - \beta_0) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{1}{4f_\xi^2(0)\bar{F}_X(d_0)}\right),$$
$$n(\hat{d}^{\ell_1} - d_0) \overset{\mathscr{L}}{\Longrightarrow} \textit{mid argmin}_{t\in\mathbb{R}} CPP(|\xi + |\alpha_0 - \beta_0|| - |\xi|, f_X(d_0)),$$

and the estimates of the parameters are asymptotically independent.

Theorem 2.3. *Consider minimizing the ℓ_2 criterion function to obtain:*

$$\hat{\theta}^{\ell_2} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha \mathbb{1}_{X_i \le d} - \beta \mathbb{1}_{X_i > d})^2$$

Then, under the assumptions of Theorem 2.1, we obtain:

$$\sqrt{n}(\hat{\alpha}^{\ell_2} - \alpha_0) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{\sigma_\xi^2}{F_X(d_0)}\right),$$
$$\sqrt{n}(\hat{\beta}^{\ell_2} - \beta_0) \overset{\mathscr{L}}{\Longrightarrow} \mathcal{N}\left(0, \frac{\sigma_\xi^2}{\bar{F}_X(d_0)}\right),$$
$$n(\hat{d}^{\ell_2} - \theta_0) \overset{\mathscr{L}}{\Longrightarrow} \textit{mid argmin}_{t\in\mathbb{R}} CPP\left(\xi + \frac{|\alpha_0 - \beta_0|}{2}, f_X(d_0)\right),$$

and the estimates of the parameters are asymptotically independent.

Observe from the above results that the asymptotic distributions of $\sqrt{n}(\hat{\alpha} - \alpha_0)$ and $\sqrt{n}(\hat{\beta} - \beta_0)$ are normal irrespective of the estimating function used for estimation, but the asymptotic variance depends upon the estimating function.
Minimax lower bound: It is evident from Theorem 2.1 - Theorem 2.3 that the rate of convergence of the change point estimator $\hat{d}$ is always n^{-1}, regardless of the tail of the error distribution and estimating function. Moreover, this rate is minimax optimal, i.e.

$$\inf_{\hat{\theta}} \sup_{P_\theta} \mathbb{E}\left[(\hat{\alpha} - \alpha_0)^2 + (\hat{\beta} - \beta_0)^2 + |\hat{d} - d_0|\right] \ge \frac{K}{n}$$

for some universal constant K, where P_θ the collection of all distribution such that X and ξ are independent and Y follows equation (2.1). This result is well-known in the literature and can be found, for example, in [22] or [33].

The more interesting part is how the asymptotic distributions of $n(\hat{d} - d_0)$ changes from the ℓ_1 to the ℓ_2 estimating function. In either case, the asymptotic distribution is characterized as the minimizer of a compound Poisson process, but the step-size is sensitive to the criterion. This has a bearing on the tail-behavior of the minimizer when ξ is heavy-tailed as articulated below in

Theorem 2.4. For notational simplicity, define F_{ℓ_i} as the limiting distribution of $n(\hat{d}-d_0)$ under the ℓ_i estimating function for $i \in \{1,2\}$:

$$F_{\ell_1}(x) = \mathbb{P}\left(\text{mid argmin}_{t\in\mathbb{R}}\text{CPP}\left(|\xi + |\alpha_0-\beta_0|| - |\xi|, f_X(d_0)\right) \le x\right) .$$
$$F_{\ell_2}(x) = \mathbb{P}\left(\text{mid argmin}_{t\in\mathbb{R}}\text{CPP}\left(\xi + \frac{|\alpha_0-\beta_0|}{2}, f_X(d_0)\right) \le x\right) .$$

As we are working with the mid argmin, both F_{ℓ_1} and F_{ℓ_2} are symmetric around 0 [e.g. see the discussion in Section 4.2 of [26]].

To compare the tail properties of F_{ℓ_1} and F_{ℓ_2} in presence of heavy tailed error, we assume the following distribution of ξ in our subsequent analysis:

$$\mathbb{P}\left(|\xi| > x\right) = \frac{1}{1+x^\gamma} \tag{2.3}$$

and ξ is symmetric around 0. This ensures that $\mathbb{E}[|\xi|^{\gamma-\nu}] < \infty$ for all $0 < \nu \le \gamma$. We next present a theorem which quantifies the tails of the asymptotic distribution of $n(\hat{d}-d_0)$ under the ℓ_1 and ℓ_2 estimating functions for the above heavy-tailed errors.

Theorem 2.4. *In our change point model equation (2.1), under the error distribution specified in equation (2.3), we have for all $x \ge k_0$:*

$$\bar{F}_{\ell_2}(x) = 1 - F_{\ell_2}(x) \ge \frac{c_0}{2f_X^\gamma(d_0)}x^{-\gamma} .$$

for some constants k_0, c_0, μ_0 explicitly mentioned in the proof. On the other hand, we have for all $x > 0$:

$$\bar{F}_{\ell_1}(x) = 1 - F_{\ell_1}(x) \le \frac{p^*}{e^{\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} - 1}\exp\left(-xf_X\left(d_0\right)\left(1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}\right)\right) .$$

where $p^* = \mathbb{P}\left(\min_{1\le i<\infty}\sum_{j=1}^{i}\left(|\xi_j + |\alpha_0-\beta_0|| - |\xi_j|\right) > 0\right) > 0.$

From Theorem 2.4, it is immediate that the asymptotic distribution of $n(\hat{d}^{\ell_2} - d_0)$ is affected by the tail index of the error distribution of ξ: it can not decay faster than $x^{-\gamma}$, whereas the asymptotic distribution of of $n(\hat{d}^{\ell_1} - d_0)$ has a sub-exponential tail[1]. Therefore for all large x, we have:

$$\mathbb{P}\left(-x \le \mathcal{D}_{\ell_2} \le x\right) \le \mathbb{P}\left(-x \le \mathcal{D}_{\ell_1} \le x\right) .$$

where $\mathcal{D}_{\ell_2}$ (resp. $\mathcal{D}_{\ell_1}$) is the limit of $n(\hat{d}^{\ell_2} - d_0)$ (resp. $n(\hat{d}^{\ell_1} - d_0)$). Therefore, it is preferable to use the change point estimator $\hat{d}^{\ell_1}$ to $\hat{d}^{\ell_2}$ for constructing an asymptotic confidence interval for all large enough levels of confidence. *To the best of our knowledge, this is the first result characterizing the tail behavior of limiting compound Poisson processes, and can be expected to be of independent interest.* More detailed empirical comparisons are presented in Section 4.

[1]Some of the constants involved in the sub-exponential tail bound of course depend on the distribution of ξ.

Proof idea: We now present a brief sketch of the proof of Theorem 2.4. In Theorems 2.2 and 2.3, we established the limiting distribution of $n(\hat{d}^{\ell_1} - d_0)$ and $n(\hat{d}^{\ell_2} - d_0)$ respectively. Both distributions are compound Poisson processes but with different step distributions: for the limit of the ℓ_2 estimator, the step distribution is $\xi + 1/2$ and for the ℓ_1 estimator, the step distribution is $|\xi + |\alpha_0 - \beta_0|| - |\xi|$. Hence, if ξ is heavy-tailed (resp. light tailed), so is the step distribution of the limit of the ℓ_2 estimator, whereas the steps of the limit of the ℓ_1 estimator are bounded (and therefore sub-gaussian) irrespective of the tail of ξ. As a compound Poisson process is closely related to the random walk corresponding to its step-size, we first establish that the tail of the minimizer of the random walk depends on that of the error distribution. In particular, in Lemmas A.2 and A.3, we show that if ξ has a power tail structure, i.e. $\mathbb{P}(|\xi| > t) \sim t^{-\gamma}$ for some $\gamma > 0$, then the tail of the minimizer of the random walk is also lower bounded by $x^{-\gamma}$. This lower bound can be translated to a lower bound on the minimizer of the compound Poisson process. On the other hand, for the limit distribution of the ℓ_1 estimator of the change point, the step distribution is sub-gaussian. Therefore, we first establish an exponential upper bound on the tail of the minimizer of a random walk with bounded steps and use it to obtain an exponential tail bound for a compound Poisson process with bounded steps. Details of the proof of Theorem 2.4 can be found in Appendix A.

Remark 2.5. *Although we have assumed a specific distribution for ξ to establish our results, an inspection of the proofs shows that the only fact essential to the calculations is the* power tail *structure of ξ, i.e. $\mathbb{P}(|\xi| > x) \sim x^{-\gamma}$ for some $\gamma > 0$. Our assumed functional form simply facilitates some routine computations and can be easily extended to the more general case. Therefore, the first conclusion of Theorem 2.4 is valid as long as ξ has power tail with index γ. We present a proof for this general tail structure in Subsection B.9 of the supplementary document. The second conclusion of Theorem 2.4 is agnostic to the tail index of ξ and continues to hold for any ξ, as long as it has finite variance. In fact, the broad conclusions of the above theorem are true for any Huber estimating function $\tilde{H}_k$ for $0 \leq k < \infty$: any such Huber function based estimate yields a sub-exponential tail for the limiting minimizer.*

Remark 2.6. *By using similar arguments to the proof of the above theorem, we can show that for a sub-gaussian ξ, both ℓ_1 and ℓ_2 criteria yield the sub-exponential concentration bound. Therefore, there is no significant gain in using robust criteria in comparison to the ℓ_2 criterion in the presence of sub-gaussian errors.*

3. Estimation in multidimensional change-problems

Summary: *This section deals with change point/plane problems in growing dimension. In Subsection 3.1, we establish that in a parallel change point estimation problem, where the number of parameters grows with the sample size, the rate of convergence of the maximal estimation error of the least square estimators is affected is by the tail of the error distribution, whereas the rate remains*

agnostic for the least absolute deviation estimator, which is further shown to be minimax optimal. In Subsection 3.2, we analyze the change plane model in growing dimension both when $p/n \to 0$ and $p \gg n$. It is also established that, the rate of convergence of the least square estimator is affected by the tail of the error distribution, whereas any estimator obtained via minimizing $\tilde{H}_k$ with $0 \le k < \infty$ achieves the minimax optimal rate regardless the tail of the error.

In the previous section, we have seen that with one-dimensional change point estimation, the advantage of using the more robust ℓ_1 estimating function is expected to confer efficiency in terms of the spread of the limiting distribution (i.e. the length of the asymptotic confidence interval), but the rate of convergence is invariant to the estimating function used. In fact, this rate can be shown to be minimax optimal, i.e. one cannot get a better rate without any further assumptions. However, the effect of using a robust estimating function is more striking when the number of change points to be estimated grows with increasing sample size.

In this section, we present two scenarios: one with many one-dimensional change points and the other with a high dimensional change-boundary, in both of which we estimate a diverging number of parameters and establish that it is possible to achieve faster rates of convergence in these situations in the presence of heavy-tailed errors using robust criteria, and in particular, the ℓ_1 criterion.

3.1. Parallel change point estimation

Suppose we have m parallel processes of one-dimensional change point models, with each process having n independent observations. Specifically, the i^{th} process carries n pairs of covariate-response pairs from the following model:

$$Y_{i,j} = \mathbb{1}_{X_{i,j} > d_{0,i}} + \xi_{i,j}\,,$$

for $1 \le j \le n$ and $1 \le i \le m$. Here, as before, we assume that $\{(X_{i,j}, \xi_{i,j})\}_{i,j}$ are i.i.d., $\xi_{i,j} \perp\!\!\!\perp X_{i,j}$ and $\xi_{i,j}$ is symmetric around 0. Furthermore, we assume that all nm pairs of observations are independent. The $\{d_{0,i}\}_{i=1}^m$'s are *free parameters* to be estimated from the data. Due to the independence among the samples, $d_{0,i}$ is estimated only from the n observations for the i^{th} problem. Define $\hat{d}_i^{\ell_1}$ and $\hat{d}_i^{\ell_2}$ to be the smallest argmin estimators obtained for the i'th problem by minimizing the ℓ_1 and ℓ_2 criteria respectively. We would like to control the estimation errors across the different problems simultaneously, hence the natural metric to consider is the maximal loss over the m problems. Specifically, we want to quantify the order of

$$\max_{1 \le i \le m} \left| \hat{d}_i^{\ell_k} - d_{0,i} \right|, \quad k = 1, 2\,.$$

We prove below that, for an appropriate growth rate of n relative to m, the maximal error only inherits the slow factor $\log m$ for the robust estimators (i.e. $\{\hat{d}_i^{\ell_1}\}_{i=1,\ldots,p}$) irrespective of the tail of the error, whereas a factor of $m^{1/\gamma}$ in unavoidable with the ℓ_2 estimates when $P(|\xi| \ge t) \sim t^{-\gamma}$.

We now present our theorem. As before, the distribution of ξ is assumed to be symmetric and $|\xi|$ is distributed as:

$$\mathbb{P}\left(|\xi| \geq t\right) = \frac{1}{1+t^{\gamma}}\,,$$

for all $t \geq 0$. Echoing Remark 2.5, the core arguments of our proof only require the power tail structure of ξ, i.e. $\mathbb{P}(|\xi| > t) \sim t^{-\gamma}$. The following theorem highlights the disparity between the rates of convergence of the maximal deviations of the ℓ_2 and ℓ_1 based estimators.

Theorem 3.1. *Suppose the change point estimator $\hat{d}_i^{\ell_2}$ for the i^{th} problem is obtained by minimizing the squared error loss. If $n/m^{1/\gamma} \to \infty$, then for any $t > 0$:*

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\leq i\leq m} \frac{n}{m^{1/\gamma}} \left|\hat{d}_i^{\ell_2} - d_{0,i}\right| > t\right) \geq c(t) > 0\,,$$

where $c(t)$ is some positive constant depending on t and other model parameters. On the other hand, if we obtain $\hat{d}_i^{\ell_1}$ by minimizing the ℓ_1 estimating function, then we have:

$$\mathbb{P}\left(\frac{n}{\log m} \max_{1\leq i\leq m} \left|\hat{d}_i^{\ell_1} - d_{0,i}\right| > t\right) \leq \frac{2e^{-c}}{1-e^{-c}} e^{-\log m\left(t\frac{f_X(d_0)}{2}(1-e^{-c})-1\right)}\,,$$

as long as $n/\log m \to 0$ for some constant c explicitly mentioned in the proof.

Proof idea: We document the main ideas behind the proof of Theorem 3.1. The first part of the theorem establishes a lower bound on the rate of convergence of the ℓ_∞ error of all estimated change points across all the problems. The main idea of the proof is that, for a finite sample of size n, the distribution of $n(\hat{d}_i^{\ell_2} - d_{i,0})$ (for any $1 \leq i \leq m$), is given by that of the minimizer of a *compound binomial process* defined as below:

$$n\left(\hat{d}_i^{\ell_2} - d_{0,i}\right) \stackrel{d}{=} \text{mid argmin}_t \sum_{k=1}^{N^i_{n,+}(t)} \left(\varepsilon_k + \frac{1}{2}\right) \mathbb{1}_{t\geq 0} + \sum_{k=1}^{N^i_{n,-}(t)} \left(\tilde{\varepsilon}_k + \frac{1}{2}\right) \mathbb{1}_{t<0}\,,$$

where the binomial processes $N_{n,+}$ and $N_{n,-}$ are defined as:

$$N^i_{n,+}(t) = \sum_{i=1}^{n} \mathbb{1}_{d_{0,i}\leq X_{i,j}\leq d_{0,i}+\frac{t}{n}} \sim \text{Bin}\left(n, F_X\left(d_{0,i}+\frac{t}{n}\right) - F_X(t)\right) \quad \forall \;\; t > 0\,,$$

$$N^i_{n,-}(t) = \sum_{i=1}^{n} \mathbb{1}_{d_{0,i}+\frac{t}{n}\leq X_{i,j}\leq d_{0,i}} \sim \text{Bin}\left(n, F_X(t) - F_X\left(d_{0,i}+\frac{t}{n}\right)\right) \quad \forall \;\; t < 0\,,$$

and then $\{\varepsilon_k\}$ are the $\xi_{i,j}$'s corresponding to the $X_{i,j}$'s satisfying $d_0 \leq X_{i,j} \leq d_0+\frac{t}{n}$ and the $\{\tilde{\varepsilon}_k\}$ are the $-\xi_{i,j}$'s corresponding to the $X_{i,j}$'s satisfying $d_0+\frac{t}{n} \leq X_{i,j} \leq d_0$.

The distribution of $n(\hat{d}_i^{\ell_2} - d_{i,0})$ is closely related to a random walk with step distribution $\xi + 1/2$, where the number of steps is derived from the binomial

processes. Therefore, we first establish a lower bound on the tail of the minimizer of the random walk and then translate that lower bound to the tail of the distribution of $n|\hat{d}_i^{\ell_2} - d_{0,i}|$ (see Lemma A.5). Finally, we use the fact for any set of independent random variables $Z_1, \ldots, Z_m$:

$$\mathbb{P}\left(\max_{1\le i\le m} Z_i > t\right) = 1 - \Pi_{i=1}^m F_{Z_i}(t) = 1 - \Pi_{i=1}^m \left(1 - \mathbb{P}(Z_i > t)\right) .$$

Hence, any lower bound on the tail of Z_i yields a lower bound on the tail of $\max_{1\le i\le m} Z_i$. Taking $Z_i = n|\hat{d}_i^{\ell_2} - d_{i,0}|$ and converting the lower bound on the tail of $n|\hat{d}_i^{\ell_2} - d_{i,0}|$ to the tail of $\max_{1\le i\le m} n|\hat{d}_i^{\ell_2} - d_{i,0}|$ concludes the first part of the proof.

The proof of the second part is similar to the first, where instead of the lower bound we establish an upper bound on the tail of the $n|\hat{d}_i^{\ell_1} - d_{0,i}|$. Note that, in case of ℓ_1 criterion:

$$n\left(\hat{d}_i^{\ell_1} - d_{0,i}\right) \stackrel{d}{=} \text{mid argmin}_t \left[\sum_{i=1}^{N_{n,+}(t)} \left(|\xi_i + 1| - |\xi_i|\right) \mathbb{1}_{t\ge 0} + \sum_{i=1}^{N_{n,-}(t)} \left(|\xi_i + 1| - |\xi_i|\right) \mathbb{1}_{t<0}\right]$$

The steps now are uniformly bounded and therefore sub-gaussian. Following the same line of arguments as in the first part of the proof, we first establish an upper bound on the tail of the minimizer of the random walk with bounded steps which is then translated to an upper bound on the tail of the $n|\hat{d}_i^{\ell_1} - d_{i,0}|$ (see Lemma A.6) and finally to the tail of $\max_{1\le i\le m} n|\hat{d}_i^{\ell_1} - d_{0,i}|$ using a union bound. The detailed proof can be found in Appendix A.

Remark 3.2. *The above theorem shows the detrimental effect of the ℓ_2 estimating funciton under heavy-tailed errors owing to the growing number of estimated parameters. The ℓ_1 based estimator is only marginally affected (by the $\log m$ factor). While we don't establish this in the paper, the HEF based estimator used in the previous section will also yield the same rate of convergence as the ℓ_1 based estimator. Further, the results are easily generalizable to the generic stump model with unknown levels on either side of the change-point with some standard technical modifications to our current proof.*

Finally we show that the rate obtained above (i.e. $n/\log m$) cannot be improved in general, even in the case of the zero error situation, i.e. this rate is minimax optimal, provided that we don't have any background information about the spread of the change points $\{d_{0,i}\}_{1\le i\le m}$.

Theorem 3.3. *Consider the above scenario of m independent change point problems where for the i'th problem the observations are generated from the following stump model:*

$$Y_{i,j} = \mathbb{1}_{X_{i,j} > d_{0,i}} + \xi_{i,j} .$$

Denote by $P_{d_{0,i}}$, the joint distribution of (X, Y) or equivalently (X, ξ) of the observations in i^{th} problem which satisfies the conditions $\xi \perp\!\!\!\perp X$ and ξ has symmetric distribution around origin. Then we have:

$$\liminf_{n\to\infty} \frac{n}{\log m} \inf_{\{\hat{d}_i\}_{1\le i\le m}} \sup_{\otimes_{i=1}^m P_{d_{i_0}}} \mathbb{E}\left[\max_{1\le i\le m} \left|\hat{d}_i - d_{i,0}\right|\right] \ge C > 0\,,$$

for some universal constant C.

3.2. Estimation of a change plane in growing dimensions

As described in Section 1, a multi-dimensional version of the canonical stump model is the so-called 'change-plane' problem:

$$Y_i = \alpha_0 \mathbb{1}_{X_i^\top d_0 \le 0} + \beta_0 \mathbb{1}_{X_i^\top d_0 > 0} + \xi_i\,. \tag{3.1}$$

where $X_i, d_0 \in \mathbb{R}^p$ and p is assumed growing with n. As d_0 is only identifiable up to its scale, we assume $d_0 \in S^{p-1}$. As before, we assume that $\{(X_i, \xi_i)\}_{i=1}^n$ are i.i.d and that ξ_i is independent of X_i with a symmetric distribution around the origin. We analyze the above canonical change plane model in two regimes: (i) when $p/n \to 0$ (Subsection 3.2.1) and (ii) when $p \gg n$ (Subsection 3.2.2). In both regimes, the dimension of d_0 is increasing with sample size, but with one fundamental difference: when $p/n \to 0$, we have many more samples than parameters and should therefore be able to estimate d_0 consistently, whereas when $p \gg n$, the problem is ill-posed and as is customary in the high dimensional literature, we need to impose a sparsity assumption on d_0: i.e. an upper bound on the number of its non-null entries. Mathematically speaking, we assume that $\|d_0\|_0 \le s$ for some unknown s which satisfies $(s \log p)/n \to 0$. Our aim is to recover the non-zero signals in d_0 consistently. We show that the rate of convergence of the change plane estimator obtained by minimizing the HEF (apart from $k = \infty$, i.e. the squared error loss) is minimax optimal in both the scenarios and is independent of the tail of the error distribution, whereas the ℓ_2 criterion based analysis (i.e. $k = \infty$) yields a slower convergence rate for heavy tailed errors, which depends on the tail index of the error.

3.2.1. When $p/n \to 0$

In the change plane estimation problem, we consider the semi-metric:

$$\begin{aligned}&\mathsf{dist}\left((\alpha_1, \beta_1, d_1), (\alpha_2, \beta_2, d_2)\right)\\ &= \sqrt{(\alpha_1 - \alpha_2)^2 + (\beta_1 - \beta_2)^2 + \mathbb{P}\left(\mathrm{sign}(X^\top d_1) \ne \mathrm{sign}(X^\top d_2)\right)}\,,\end{aligned}$$

which is motivated by the one used by [23] (see Chapter 14), with the only difference being that instead of considering the Euclidean distance between two candidate change-plane vectors d_1 and d_2 we use the mass of the wedge bounded

by the two corresponding corresponding hyperplanes to define a metric. This particular metric is geometrically convenient to analyze in the change-plane problem as will be seen in our subsequent computations and can be easily related to the ℓ_2 distance under an additional condition which is satisfied under various distributional assumptions on the covariate X.

Define $\theta = (\alpha, \beta, d)$. In the growing dimension regime, the rates of convergence of the estimates are affected by the underlying dimension. We show later in this section (see Theorem 3.6) that for HEF with $0 \leq k < \infty$ (i.e. excluding squared error loss), the corresponding Huber estimator satisfies:

$$\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1} \mathsf{dist}^2\left(\hat{\theta}, \theta_0\right) = O_p(1)\,. \tag{3.2}$$

The above rate can be converted to a rate of convergence of the ℓ_2 estimation error $\|\hat{d} - d_0\|_2$ of the change plane parameter via Assumption 3.5 stated below. In contrast, the rates for the least squares estimators are found to be slower and are non-trivially affected by the tail of the error distribution. Although the rate in equation (3.2) is shown to be minimax optimal for the change plane estimator (Theorem 3.7), the rate of convergence of the one dimensional parameters (α_0, β_0) can be further boosted to $\sqrt{n}$ provided that the estimation error of the change plane estimator is smaller than $n^{-1/2}$ (i.e. $p \leq \sqrt{n}/\log(n/p)$)using the following two step procedure:

1. Get initial estimates of (α_0, β_0) and an estimate of d_0 as follows:

$$(\hat{\alpha}^k_{\text{init}}, \hat{\beta}^k_{\text{init}}, \hat{d}^k) = \underset{(\alpha>\beta)\in\Omega, d\in S^{p-1}}{\operatorname{argmin}} \sum_i \tilde{H}_k\left(Y_i - \alpha \mathbb{1}_{X_i^\top d \leq 0} - \beta \mathbb{1}_{X_i^\top d > 0}\right).$$

2. Update the estimates of α_0, β_0 obtained in the previous step as follows:

$$(\hat{\alpha}^k, \hat{\beta}^k) = \underset{(\alpha>\beta)\in\Omega}{\operatorname{argmin}} \sum_i \tilde{H}_k\left(Y_i - \alpha \mathbb{1}_{X_i^\top \hat{d}^k \leq 0} - \beta \mathbb{1}_{X_i^\top \hat{d}^k > 0}\right).$$

where as before, we assume that $(\alpha_0, \beta_0) \in \Omega$ for a compact subset $\Omega \subseteq \mathbb{R}^2$ for technical simplicity. The assumption $\alpha_0 > \beta_0$ is for identifiably as one can reverse their order simply by changing the sign of d. The intuition for this rate acceleration is the following: if $\hat{d}^k$ converges to d_0 at a faster rate than $\sqrt{n}$, then we can re-estimate (α_0, β_0) at $\sqrt{n}$ - rate from the following surrogate model:

$$Y_i = \alpha_0 \mathbb{1}_{X_i^\top \hat{d}^k \leq 0} + \beta_0 \mathbb{1}_{X_i^\top \hat{d}^k > 0} + \xi_i\,,$$

where we simply replace d_0 by its estimate $\hat{d}^k$. If the estimation error of $\hat{d}^k$ is larger than $n^{-1/2}$, it is not possible to recover the parametric convergence rate for estimates of (α_0, β_0).

We now state our assumptions and the theorems.:

Assumption 3.4. *Our parameter space Ω for (α, β) is a compact subset of $\mathbb{R}^2$ such that for any $(\alpha, \beta) \in \Omega$, $\alpha > \beta$. The hyperplane parameter $d_0 \in S^{p-1}$.*

Our next assumption (henceforth referred as *wedge assumption*) relates the probability of X lying in between two hyperplanes to the angle between those two hyperplanes.

Assumption 3.5. *We assume there exists some $\delta > 0$ such that:*

$$\mathbb{P}\left(sign(X^\top d) \neq sign(X^\top d_0)\right) \geq c\|d - d_0\|_2$$
$$\mathbb{P}\left(X^\top d \wedge X^\top d_0 \geq 0\right) \geq C_1$$
$$\mathbb{P}\left(X^\top d \vee X^\top d_0 \leq 0\right) \geq C_2$$

for all $\|d - d_0\|_2 \leq \delta$, where the constants c, C_1, C_2, δ do not depend on n.

The first condition can be interpreted as saying that if we choose two hyperplanes $X^\top d = 0$ and $X^\top d_0 = 0$ the probability of X falling in between these hyperplanes is bounded below, up to a constant, by the angle between the hyperplanes. This assumption can be thought as an analogue of the restricted eigenvalue assumption frequently used in the analysis of the high dimensional linear model (especially LASSO, see e.g. [6]) to obtain the estimation error from the prediction error and was also used in earlier work by the authors [31], where it was shown that the condition is satisfied by several classes of distributions (e.g. under elliptical symmetry (Lemma B.1), log-concavity of densities (Lemma C.9)). The second and third inequalities are weak assumptions, which ensure that the support of X is not restricted to the one side of the hyperplane.

We next state our theorems for this regime:

Theorem 3.6 (Rate of convergence)**.** *Suppose we estimate $\theta_0 = (\alpha_0, \beta_0, d_0)$ using the two-shot approach described above, i.e. by minimizing the scaled HEF and then re-estimating (α_0, β_0). Then, under Assumptions 3.4-3.5, we have for $0 \leq k < \infty$:*

$$\left(\sqrt{n} \wedge \frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\right)\left(\hat{\alpha}^k - \alpha_0\right) = O_p(1),$$
$$\left(\sqrt{n} \wedge \frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\right)\left(\hat{\beta}^k - \beta_0\right) = O_p(1),$$
$$\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\mathbb{P}\left(sign(X^\top \hat{d}^k) \neq sign(X^\top d_0)\right) = O_p(1)\,,$$

which along with Assumption 3.5 yields:

$$\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\left\|\hat{d}^k - d_0\right\|_2 = O_p(1)\,.$$

For $k = \infty$, i.e. under squared error loss we have under further assumption $\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] < \infty$:

$$\left(\sqrt{n} \wedge \frac{n}{p\,\|\xi\|_{n,L_1}}\left(\log\frac{n}{p\,\|\xi\|_{n,L_1}}\right)^{-1}\right)\left(\hat{\alpha}^{\ell_2} - \alpha_0\right) = O_p(1)$$

$$\left(\sqrt{n} \wedge \frac{n}{p\,\|\xi\|_{n,L_1}} \left(\log \frac{n}{p\,\|\xi\|_{n,L_1}}\right)^{-1}\right) \left(\hat{\beta}^{\ell_2} - \beta_0\right) = O_p(1)$$

$$\frac{n}{p\,\|\xi\|_{n,L_1}} \left(\log \frac{n}{p\,\|\xi\|_{n,L_1}}\right)^{-1} \left\|\hat{d}^{\ell_2} - d_0\right\|_2 = O_p(1)\,.$$

where $\|\xi\|_{n,L_1} = \mathbb{E}\left[\max_{1\le i\le n} |\xi_i|\right]$.

Like the results of the previous subsection, Theorem 3.6 shows that in a growing dimension setting the rate of convergence of the Huber estimator for any $0 \le k < \infty$ is faster than using the standard squared error loss: the rate of the least squares estimator of d_0 suffers from an additional factor $\|\xi\|_{n,L_1}$, which depends on the tail of the distribution of ξ. We note that this is not that an isolated phenomenon, e.g. in non-parametric regression, the rate of convergence of the least square estimators is similarly affected by the tail of the error, e.g. see [17]. Note that in the fixed p regime this factor can be ignored via a different maximal inequality which, when used in growing dimensional regime, yields the rate n/p^2. More specifically, consider Lemma 2.14.1 of [38], which we state here for the ease of our readers:

$$\mathbb{E}\left[\sqrt{n} \sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)f|\right] \lesssim \mathbb{E}\left[\mathcal{J}(\theta_n, \mathcal{F})\|F\|_{2,n}\right] \lesssim \mathcal{J}(1,\mathcal{F})\sqrt{\mathbb{E}[F^2]}\,,$$

where F is the envelope of $\mathcal{F}$ and $\mathcal{J}$ quantifies the complexity of $\mathcal{F}$ as follows:

$$\mathcal{J}(\delta,\mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))}\; d\epsilon\,.$$

and $\theta_n = \sup_{f\in\mathcal{F}} \|f/F\|_{2,n}$ (with the convention $0/0 = 0$). The rate of convergence of the least squares estimator in Theorem 3.6 is obtained via a modified version of the first inequality (details can be found in the proof), whereas one may also use the weaker second inequality which, in this case, yields the rate (n/p^2). Combining this with the rate obtained in Theorem 3.6 leads to the following modified rate of convergence:

$$\left\|\hat{d}^{\ell_2} - d_0\right\|_2 = O_p\left(\frac{p\,\|\xi\|_{n,L_1}}{n} \log \frac{n}{p\,\|\xi\|_{n,L_1}} \wedge \frac{p^2}{n}\right)\,.$$

When p is fixed, the second weaker inequality yields a better rate of convergence as the factor p^2 is a constant. In the growing dimension regime, the VC dimension of the underlying function class is growing with the sample size, hence the interplay between the ambient dimension p and the tail of the error distribution starts affecting the rate of convergence of the least squares estimator. However, the tail factor $\|\xi\|_{n,L_1}$ does not appear in the rate of the other Huber estimators, as the criterion function becomes bounded irrespective of the thickness of tail of the distribution.

To summarize, we have established in Section 2 that the robust Huber estimators (for any $0 \leq k < \infty$) yield a more concentrated limiting distribution than the squared error loss, whereas in the growing dimension regime, the effect is more prominent: robust Huber estimators yield a faster rate of convergence, which is also minimax optimal as shown in our next theorem. This underscores the necessity of using robust estimators in high dimensional change plane problems, especially in presence of heavy tailed errors.

Theorem 3.7 (Minimax lower bound)**.** *Suppose* $\mathcal{P} = \left\{P_d : d \in S^{p-1}\right\}$ *is the collection of all change plane models such that the distribution* P_d *of* (X, Y) *or equivalently the distribution of* (X, ξ) *satisfies the following:*

$$Y = \mathbb{1}_{X^\top d > 0} + \xi$$

where X *is independent of* ξ *and* ξ *is symmetric around the origin. Then we have:*

$$\inf_{\hat{d}} \sup_{P_d} \mathbb{E}_\theta \left(\mathsf{dist}^2(\hat{d}, d)\right) \geq K \frac{p}{n} \left(1 + \log \frac{p}{n}\right),$$

where K *is a universal constant and the semi-metric* dist *is defined as:*

$$\mathsf{dist}(d_1, d_2) = \sqrt{\mathbb{P}\left(sign(X^\top d_1) \neq sign(X^\top d_2)\right)}.$$

Hence, the change plane estimator obtained in Theorem 3.6 via the Huber estimating equation $\tilde{H}_k$ *for* $0 \leq k < \infty$ *(i.e. excluding squared error loss) is minimax optimal.*

Remark 3.8. *Notice that we restrict our minimax calculation only to the change plane parameter* d_0 *assuming we know* (α_0, β_0) *(in fact, without loss of generality we assume* $\alpha_0 = 0, \beta_0 = 1$*), as the minimaxity of the rate of convergence of* (α_0, β_0) *is immediate and not interesting. Theorem 3.7 indicates that any Huber estimator for* $0 \leq k < \infty$ *is minimax optimal. The proof of this theorem relies on a clever construction of the local alternatives and an application of Fano's inequality (e.g. see [42]).*

3.2.2. When $p \gg n$

In this section, we present our analysis of the change plane estimator in the regime $p \gg n$, i.e. the HDLSS (high dimension low sample size) setting. As is true for any high dimensional model, consistent estimate of d_0 is not information theoretically possible without further restrictions on the parameter space. A typical condition frequently imposed on the parameter space is that of sparsity: there exists some (unknown) s such that only s many elements of d_0 are non-zero, where s may also increase with the sample size. We summarize this in the following assumption:

Assumption 3.9. *The true change plane direction* d_0 *is sparse, i.e. there exists* s *such that* $\|d_0\|_0 \leq s$ *where* s *may slowly grow with* n*, satisfying* $(s \log p)/n \to 0$.

To estimate d_0 under this sparsity constraint, we follow the *structural risk minimization* method, an idea originally from [39] and later implemented in a series of work (e.g. [29], [3], [4] and references therein). The key idea is to use a penalty function to balance the bias-variance trade-off. To understand this, consider our stump model:

$$Y_i = \alpha_0 \mathbb{1}_{X_i^\top d_0 \leq 0} + \beta_0 \mathbb{1}_{X_i^\top d_0 > 0} + \xi_i \,.$$

Define $\mathcal{F}_m$ to be set of all hyperplanes with sparsity at-most m, i.e.:

$$\mathcal{F}_m = \left\{ f_\theta(X) = \alpha \mathbb{1}_{X^\top d \leq 0} + \beta \mathbb{1}_{X^\top d > 0} : (\alpha, \beta) \in \Omega, \|d\|_0 \leq m \right\} ,$$

for $1 \leq m \leq p$, where as before, we denote by $\theta = (\alpha, \beta, d)$, the collection of all the parameters. Now for each m, we define the empirical minimizer $\hat{\theta}_m^k := \hat{\theta}_{m,n}^k$ as:

$$\begin{aligned}
\hat{\theta}_m^k &= \underset{\theta : f_\theta \in \mathcal{F}_m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \tilde{H}_k \left(Y_i - f_\theta(X_i) \right) \\
&= \underset{\theta : f_\theta \in \mathcal{F}_m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left[\tilde{H}_k \left(Y_i - f_\theta(X_i) \right) - \tilde{H}_k \left(Y_i - f_{\theta_0}(X_i) \right) \right] \\
&= \underset{\theta : f_\theta \in \mathcal{F}_m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left[\tilde{H}_k \left(Y_i - f_\theta(X_i) \right) - \tilde{H}_k \left(\xi_i \right) \right]
\end{aligned}$$

for $0 \leq k \leq \infty$. The corresponding population minimizer is defined as:

$$\theta_m^k = \underset{d : f_d \in \mathcal{F}_m}{\operatorname{argmin}} \mathbb{E} \left[\tilde{H}_k \left(Y - f_\theta(X) \right) - \tilde{H}_k \left(\xi \right) \right] .$$

Note that, the larger the m, the more complex is the function class $\mathcal{F}_m$ (as $\mathcal{F}_{m_1} \subseteq \mathcal{F}_{m_2}$ for any $m_1 \leq m_2$), and consequently, the variance starts dominating the bias for large values of m. In other words, $\hat{\theta}_m^k$ has smaller training error, but larger generalization error for large values of m. Therefore, to choose an optimal model m, we add a penalty $\operatorname{pen}(m)$ (which quantifies the complexity of the model $\mathcal{F}_m$ and is increasing in m) to the training error of $\hat{\theta}_m^k$ and choose the one which minimizes the penalized training error:

$$\hat{m}^k = \underset{1 \leq m \leq p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left[\tilde{H}_k \left(Y_i - f_{\hat{\theta}_m^k}(X_i) \right) - \tilde{H}_k(\xi_i) \right] + \operatorname{pen}(m) \,.$$

and set the final estimator as $\hat{\theta}_{\hat{m}^k}^k$. The penalty function should be chosen carefully depending on the complexity of the underlying function class to balance the bias-variance tradeoff. We quantify the complexity of $\mathcal{F}_m$ using its VC dimension. It follows from Lemma 1 of [1] that the VC dimension of $\mathcal{F}_m$ is,

$$V_m = VC \left(\mathcal{F}_m \right) \asymp m \log \frac{ep}{m} \,. \tag{3.3}$$

Based on the above notion of complexity, we use the following penalty function:

$$\text{pen}(m) = \kappa \left(\frac{V_m \log (n/V_m)}{n} \right) . \tag{3.4}$$

for some constant κ independent of n, while using HEF $\tilde{H}_k$ for $0 \leq k < \infty$. For $k = \infty$, i.e. while using the least squares estimator, we use a slightly different penalty (see Theorem 3.14 for more details). Therefore, our pen function is the VC dimension of the model under consideration (up to a constant and a log factor). Finally, we can accelerate the rate of convergence of (α_0, β_0) by following the same procedure as prescribed in Subsection 3.2.1): i.e., first estimate d_0 by minimizing the penalized criterion function, then re-estimate (α_0, β_0) using $\hat{d}^k_{\hat{m}^k}$ as a proxy for d_0. Henceforth, we denote by $(\hat{\alpha}^k, \hat{\beta}^k, \hat{d}_{\hat{m}^k})$ as these final estimators obtained via the two-shot procedure.

Our next theorem presents the rate of convergence of the above estimates for $0 \leq k < \infty$:

Theorem 3.10. *Under Assumptions 3.4, 3.5 and 3.9 and using the penalty introduced in (3.4) we have:*

$$\left(\sqrt{n} \wedge \frac{n}{V_s \log \frac{n}{V_s}} \right) \left(\hat{\alpha}^k - \alpha_0 \right) = O_p(1),$$
$$\left(\sqrt{n} \wedge \frac{n}{V_s \log \frac{n}{V_s}} \right) \left(\hat{\beta}^k - \beta_0 \right) = O_p(1),$$
$$\frac{n}{V_s \log \frac{n}{V_s}} \left\| \hat{d}^k_{\hat{m}^k} - d_0 \right\|_2 = O_p(1) .$$

Remark 3.11. *From equation (3.3), it is readily seen that the rate of convergence of the change plane estimator $\hat{d}$ is:*

$$\left\| \hat{d}^k_{\hat{m}^k} - d_0 \right\|_2 = O_p \left(\frac{s \log (ep/s)}{n} \log \left(\frac{n}{s \log (ep/s)} \right) \right)$$

i.e. upto a log factor, the rate if $s \log p/n$, which can be thought as the high dimensional analogue of $1/n$ (the rate obtained for the change point estimator in finite dimension) in presence of sparsity.

We now present our results regarding the minimax lower bound for this change plane problem in this HDLSS scenario under the sparsity constraint. As before, we restrict our attention to the parameter of interest d_0 and assume we know α_0, β_0, in particular setting $\alpha_0 = 0, \beta_0 = 1$.

Theorem 3.12. *Assume $\mathcal{P} = \left\{ P_d : d \in S_s^{p-1} \right\}$ is the collection of all change plane models with S_s^{p-1} being the set of all unit vectors in dimension p with sparsity at-most s, where the distribution P_d of (X, Y) or equivalently the distribution of (X, ξ) satisfies the following:*

$$Y = \mathbb{1}_{X^\top d > 0} + \xi ,$$

where X is independent of ξ and ξ is symmetric around the origin. Then we have:

$$\inf_{\hat{d}} \sup_{P_d} \mathbb{E}_\theta \left(\|\hat{d} - d_0\|^2 \right) \geq K \left(\frac{s \log (ep/s)}{n} \right)^2 .$$

Squared error loss: The rate of convergence of the least squares estimator for this regime is also compromised by the tail of the error distribution, which is in agreement with our findings in Subsection 3.2.1. To establish the theoretical properties of the LSE, we slightly strengthen our sparsity assumption below:

Assumption 3.13. *The true change plane direction d_0 is sparse, i.e. there exists s such that $\|d_0\|_0 \leq s$ where s may slowly grow with n, satisfying*

$$\frac{s(\log p)^{(1+\delta)} \|\xi\|_{n,2}}{n} \to 0$$

where $\|\xi\|_{n,2} = \sqrt{\mathbb{E}[\max_{1 \leq i \leq n} \xi_i^2]}$, which is assumed to be finite.

Two comments on this modified sparsity assumption are in order: first note that, we need a slightly higher power of $\log p$ in comparison to its counterpart in Assumption 3.9. This is likely a technical artifact and possibly avoidable with more tedious analysis. Next, we also have an additional term $\|\xi\|_{n,2}$ which captures the effect of the tail of the error distribution in the rate of the LSE, similar to what we see in Theorem 3.6. This modified assumption necessitates changing our penalty to:

$$\text{pen}(m) = \frac{V_m (\log p)^\delta \|\xi\|_{n,2}}{n} \log \frac{n}{V_m} \tag{3.5}$$

where, as before, V_m is the VC dimension of $\mathcal{F}_m$. The following theorem establishes the rate of convergence of the LSE.

Theorem 3.14. *Suppose we estimate $\theta_0 = (\alpha_0, \beta_0, d_0)$ using the two-shot procedure under squared error loss. Then under Assumptions 3.4, 3.5 and 3.13 we obtain:*

$$\left(\sqrt{n} \wedge \frac{n}{s(\log p)^{(1+\delta)} \|\xi\|_{n,2}} \left(\log \frac{n}{s \log p} \right)^{-1} \right) (\hat{\alpha} - \alpha_0) = O_p(1),$$

$$\left(\sqrt{n} \wedge \frac{n}{s(\log p)^{(1+\delta)} \|\xi\|_{n,2}} \left(\log \frac{n}{s \log p} \right)^{-1} \right) \left(\hat{\beta} - \beta_0 \right) = O_p(1),$$

$$\frac{n}{s(\log p)^{(1+\delta)} \|\xi\|_{n,2}} \left(\log \frac{n}{s \log p} \right)^{-1} \left\| \hat{d}^{\ell_2}_{\hat{m}^k} - d_0 \right\|_2 = O_p(1) .$$

Remark 3.15. *A remark similar to Remark 3.8 is in order: Theorem 3.12 conveys a similar message as Theorem 3.7, i.e. any Huber-estimator for $0 \leq k < \infty$ is minimax optimal up to a log factor, whereas the least squares estimator is not (as seen above), especially when the distribution of ξ has a heavy tail. Therefore, as in the previous subsection, robust Huber-estimators are preferable to the least squares estimator in this high dimensional regime.*

Remark 3.16. *In this paper we only dealt with the cases where $p/n \to 0$ and $p \gg n$, not when $p/n \to \alpha \in (0, \infty)$. The primary reason why the latter requires different techniques is as follows: analysis of both the regimes $p/n \to 0$ and $p \gg n$ are similar to fixed dimensional problem to some extent. When $p/n \to 0$, although the dimension is growing, but is much less than the sample size, permitting the number of observations per co-ordinate of the underlying parameter (here d_0) goes to ∞, which leads us to use similar techniques employed to deal with finite dimensional problem, but with some important technical changes to take care of the growing dimension. Similarly for $p \gg n$, a typical assumption is that of sparsity, which essentially dictates that the number of non-zero co-ordinates of d_0 is much less than the sample size. The extra difficulty here is to detect which parameters are non-zero, for which we use a penalty based on the complexity of the underlying function class (e.g. ℓ_1 penalty in LASSO) to prevent overfitting and consequently identify the support of the parameter. In a nutshell, the number samples per co-ordinate of the underlying parameter to be estimated diverges in both the regimes.*

However, when $p < n$ and $p/n \to \alpha$ (say with $\alpha = 1/2$), then there is a-priori no reason to assume sparsity and therefore the number of effective sample per coordinate of the the unknown parameter to be estimated is 2 asymptotically. In this case, all traditional statistical analysis for finite dimension or for high dimension with sparsity assumption fails and often either a bias creeps in, or variance of the asymptotic distribution is inflated. Problems in this regime are extremely hard to analyze, require a completely different set of tools (e.g. AMP introduced in [12] or the techniques used in [14]) and till date the regime has been investigated only for the linear regression model and certain kinds of GLMS, albeit under strong assumptions (e.g. gaussianity or subgaussianity). We believe that a non-standard problem like the canonical change plane estimator in this regime is currently insolvable in the p/n converging to a constant regime.

4. An empirical study of the quantiles of the limiting distributions

In this section we present tables of quantiles of the limit distributions of change point estimator under both the ℓ_1 and ℓ_2 criteria. In Section 2, we established theoretically (Theorem 2.2 and 2.3) that in the presence of heavy tailed errors, the limiting distribution of the change point estimator under ℓ_1 criterion has a thinner tail (i.e. more concentrated asymptotic confidence interval) than the change point estimator under ℓ_2 criterion. We provide some illustrations of this phenomenon in our simulations below.

We generate data from the following stump model:

$$Y_i = \mu \mathbb{1}_{X_i \geq d_0} + \xi_i \,.$$

where we have assumed $d_0 = 0$, $X_i \sim \mathcal{N}(0, 1)$. Recall that the limiting distribution of $\hat{d}^{\ell_1}$ is (see Theorem 2.2):

$$n(\hat{d}^{\ell_1} - d_0) \overset{\mathscr{L}}{\Longrightarrow} \text{mid argmin}_{t \in \mathbb{R}} \text{CPP}\left(|\xi + \mu| - |\xi|, f_X(d_0)\right)$$

and the limiting distribution of $\hat{d}^{\ell_2}$ is (see Theorem 2.3):

$$n(\hat{d}^{\ell_2} - d_0) \overset{\mathscr{L}}{\Longrightarrow} \text{mid argmin}_{t\in\mathbb{R}}\text{CPP}\left(\xi + \frac{\mu}{2}, f_X(d_0)\right)$$

For ξ, we consider seven different distributions: standardized $T_3, T_4, T_5, T_6, T_{10}$, T_{15} (i.e. $\mathsf{var} = 1$) and $\mathcal{N}(0,1)$, while for the signal μ we consider four different values $\mu = 0.1, 0.5, 1, 2$. We present here 8 different tables: two tables (ℓ_1 and ℓ_2 quantiles) for each value of μ. Each table consists of five different one sided quantiles $(90\%, 95\%, 97.5\%, 99\%, 99.5\%)$ for each of the five different distributions of ξ (calculated using 10^6 monte-carlo iterations). Recall that as we compute the mid argminchange-point estimator, the limit distributions are symmetric and it suffices to report the upper quantiles. The percentages presented inside the brackets following the quantiles in the even-numbered tables show the relative change in the ℓ_2 based quantile as compared to the ℓ_1 based counterpart.

TABLE 1
Quantiles of asymptotic distribution under ℓ_1 criterion using $\mu = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	717.0	1152.3	1600.1	2100.3	2324.8
T_4	943.7	1470.5	1924.0	2308.9	2432.3
T_5	1062.0	1611.0	2045.6	2366.8	2460.7
T_6	1133.6	1690.1	2110.6	2389.5	2475.5
T_{10}	1247.8	1808.5	2196.3	2419.7	2489.5
T_{15}	1294.4	1859.2	2229.8	2433.0	2499.5
Normal	1381.6	1944.1	2278.8	2449.6	2509.1

TABLE 2
Quantiles of asymptotic distribution under ℓ_2 criterion using $\mu = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1045.7(+45.8%)	1584.7(+37.5%)	2029.7(+26.8%)	2358.3(+12.3%)	2457.6(+5.7%)
T_4	1050.1(+11.3%)	1598.2(+8.7%)	2034.3(+5.7%)	2355.6(+2%)	2456.3(+1%)
T_5	1055.6(-0.6%)	1600.9(-0.6%)	2040.0(-0.3%)	2364.6(-0.1%)	2461.0(+0.01%)
T_6	1056.2(-5.1%)	1601.2(-5.3%)	2043.0(-3.2%)	2366.0(-1%)	2461.1(-0.6%)
T_{10}	1052.5(-15.6%)	1593.9(-11.9%)	2038.0(-7.2%)	2363.0(-2.3%)	2460.05(-1.2%)
T_{15}	1054.8(-18.5%)	1601.0(-13.9%)	2046.2(-8.2%)	2366.1(-2.75%)	2461.0(-1.5%)
Normal	1051.4(-24%)	1600.9(-17.6%)	2044.5(-10.3%)	2363.9(-3.49%)	2460.0(-2%)

TABLE 3
Quantiles of asymptotic distribution under ℓ_1 criterion using $\mu = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	28.5	46.8	67.7	98.1	122.5
T_4	38.2	62.7	89.9	129.7	162.2
T_5	44.5	73.1	104.8	150.7	188.1
T_6	48.3	78.9	113.5	163.3	203.6
T_{10}	55.7	90.7	130.6	187.4	233.4
T_{15}	59.0	97.0	139.7	200.5	250.0
Normal	66.1	108.3	155.7	224.8	279.9

TABLE 4
Quantiles of asymptotic distribution under ℓ_2 criterion using $\mu = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	45.8(+60.7%)	77.3(+65.2%)	113.5(+67.6%)	166.9(+70.1%)	207.7(69.5%)
T_4	47.3(+23.8%)	78.3(+24.9%)	113.3(+26%)	163.5(+26.1%)	204.5 (+26.1%)
T_5	48.1(+8.1%)	78.9(+7.9%)	113.0(+7.8%)	163.3(+8.4%)	202.7(+7.8%)
T_6	48.3(+0%)	79.0(+0.1%)	113.5(+0%)	162.8(-0.3%)	203.1(-0.25%)
T_{10}	48.6(-12.7%)	79.1(-12.8%)	113.3(-13.2%)	162.6(-13.2%)	202.4(-13.3%)
T_{15}	48.8(-17.3%)	79.3(-18.2%)	113.7(-18.6%)	162.8(-18.8%)	203.2(-18.72%)
Normal	48.8(-26.2%)	79.4(-26.7%)	113.8(-27%)	163.0(-27.5%)	202.7(-27.6%)

TABLE 5
Quantiles of asymptotic distribution under ℓ_1 criterion using $\mu = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	8.4	13.5	19.2	27.6	34.4
T_4	10.5	16.9	24.0	34.3	42.7
T_5	11.7	18.7	26.8	38.2	47.5
T_6	12.5	20.2	28.9	41.2	51.3
T_{10}	14	22.7	32.4	46.4	57.8
T_{15}	14.7	23.9	34.1	48.9	60.9
Normal	16.1	26.2	37.2	53.7	66.7

TABLE 6
Quantiles of asymptotic distribution under ℓ_2 criterion using $\mu = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	11.8(+40.5%)	20.3(+50.4%)	30.7(+59.9%)	46.1(+67.0%)	59.3(+72.4%)
T_4	12.7(+21%)	21.0(+24.3%)	30.5(+27.1%)	44.6(+30.1%)	56.0(+31.1%)
T_5	13.0(+11.1%)	21.3(+14%)	30.5(+13.8%)	44.0(+15.2%)	55.0(+15.8%)
T_6	13.1(+4.8%)	21.4(+5.9%)	30.6(+5.9%)	44.0(+6.8%)	55.0(+7.2%)
T_{10}	13.4(-4.28%)	21.5(-5.29%)	30.7(-5.25%)	43.8(-5.6%)	54.1(-6.4%)
T_{15}	13.4(-8.8%)	21.6(-9.6%)	30.7(-10%)	43.9(-10.2%)	54.0(-11.3%)
Normal	13.5(-16.1%)	21.7(-17.2%)	30.6(-17.1%)	43.5(-19%)	54.0(-19%)

TABLE 7
Quantiles of asymptotic distribution under ℓ_1 criterion using $\mu = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1.7	4.8	7.5	11.1	13.8
T_4	2.7	5.7	8.5	12.3	15.3
T_5	3.1	6.1	9.0	13.0	16.0
T_6	3.3	6.3	9.3	13.2	16.4
T_{10}	3.6	6.7	9.8	14.0	17.2
T_{15}	3.8	6.9	10.0	14.3	17.6
Normal	4.0	7.2	10.3	14.7	18.1

TABLE 8
Quantiles of asymptotic distribution under ℓ_2 criterion using $\mu = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	2.2(+29.4%)	5.9(+23%)	9.5(+26.7%)	15.0(+35.1%)	19.8(+43.5%)
T_4	2.9(+7.4%)	6.2(+8.8%)	10.0(+17.6%)	14.3(+16.3%)	18.2(+19%)
T_5	3.2(+3.2%)	6.4(+3.3%)	9.6(+6.7%)	14.2(+9.2%)	17.8(+11.25%)
T_6	3.3(+0%)	6.4(+1.6%)	9.6(+3.2%)	14.0(+6.1%)	17.5(+6.7%)
T_{10}	3.5(-2.8%)	6.6(-1.5%)	9.7(-1%)	14.0(+0%)	17.4(+1.2%)
T_{15}	3.6(-5.3%)	6.6(-4.3%)	9.7(-3%)	13.9(-2.8%)	17.3(-1.7%)
Normal	3.7(-7.5%)	6.7(-7%)	9.7(-5.8%)	13.9(-5.4%)	17.1(-5.5%)

From the above tables, it is immediate that if the error distribution has heavy tails, say, T_3, T_4, it is preferable to use $\hat{d}^{\ell_1}$ to $\hat{d}^{\ell_2}$ as the former has tighter limiting confidence interval for any of the levels presented in our tables. On the other hand, if the error distribution is normal, then the ℓ_2 estimator is more efficient in terms of the width of the asymptotic confidence interval, as it is maximum likelihood estimator of d_0. In fact, the ℓ_2 estimator starts becoming efficient for T distributions with higher degrees of freedom as is already evident from the above tables where we see a reduction is some of the ℓ_2 quantiles for certain values of μ with T_6 error, and a systematic reduction with T_{10} and T_{15} errors.

Remark 4.1. *We have added a section (Section C) to the supplementary document, where we have extended the experiments for several intermediate values of* $k \in \{0.1, 0.5, 1, 2, 5, 10\}$ *and for the same four signals and same seven distributions as in this main paper. The intermediate values of* k *exhibit an expected monotone trend with respect to quantile behavior: for heavier tailed distributions e.g.* T_3, T_4*, smaller values of* k *correspond to smaller quantiles (narrower confidence regions), whilst, for the light-tailed normal distribution, the quantiles decrease with increasing* k*.*

5. Conclusion

In this paper, we have analyzed various estimators in the standard change point model and its multi-dimensional analogue by minimizing HEFs, especially in the presence of heavy tailed errors. We note that the robust Huber-estimators show varying degrees of advantage over the least squares estimator, depending on the dimensionality of the problem.

1. In one dimension, all estimators achieve the same rate of convergence, whereas the limiting distributions for the robust criteria based estimators are more concentrated around 0 than that of the least squares estimator. This effect diminishes as the tail of the error distribution becomes lighter: in particular, for normal errors the least squares estimator has a narrower asymptotic confidence interval in comparison to the robust estimators. We believe a similar phenomenon will arise in the change-plane problem for fixed p (where again, all the Huber estimators and the LSE will converge at rate n), but the limit distributions in the multidimensional case are expected to be multidimensional analogues of compound Poisson processes with extremely involved characterizations. Almost nothing is known about these objects and their study constitutes a highly non-trivial project in its own right.
2. In growing dimensions, the robust estimators attain faster rates of convergence than the least squares estimator, in particular attaining the minimax rate which does not depend upon the tail of the error, whilst the rate of convergence of the least squares estimator is dampened by the tail of the error distribution.

We now briefly discuss some variants of the problem considered above as well as possible directions for future research.

5.1. *Binary response model:*

A natural variant of the change plane model analyzed in Subsection 3.2 is the following binary response model:

$$X \sim P, \quad \mathbb{P}(Y = 1 \mid X) = \alpha_0 \mathbb{1}_{X^\top d_0 \le 0} + \beta_0 \mathbb{1}_{X^\top d_0 > 0}\,,$$

where $0 < \alpha_0 \neq \beta_0 < 1$. One may minimize the squared error loss to estimate the unknown parameters:

$$(\hat{\alpha}, \hat{\beta}, \hat{d}) = \underset{\alpha,\beta,d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \alpha \mathbb{1}_{X_i^\top d \le 0} - \beta \mathbb{1}_{X_i^\top d > 0} \right)^2 = \underset{\alpha,\beta,d}{\operatorname{argmin}} \mathbb{P}_n f_{\alpha,\beta,d}\,.$$

As in Subsection 3.2, the change plane parameter d_0 here is also identified up to its direction and the level parameters (α_0, β_0) are identified up to their order, so we assume $\|d_0\| = 1$ and $\alpha_0 < \beta_0$. The loss function $f_{\alpha,\beta,d}$ is uniformly bounded by 1, hence the techniques used to prove the first part of Theorem 3.6 yield:

$$\left(\sqrt{n} \wedge \frac{n}{p} \left(\log \frac{n}{p} \right)^{-1} \right) (\hat{\alpha} - \alpha_0) = O_p(1)$$

$$\left(\sqrt{n} \wedge \frac{n}{p} \left(\log \frac{n}{p} \right)^{-1} \right) \left(\hat{\beta} - \beta_0 \right) = O_p(1),$$

$$\frac{n}{p} \left(\log \frac{n}{p} \right)^{-1} \mathbb{P} \left(\operatorname{sign}(X^\top \hat{d}) \neq \operatorname{sign}(X^\top d_0) \right) = O_p(1)\,.$$

Furthermore, the rate obtained above can be shown to be minimax optimal (up to a log factor) by following a similar line argument as in the proof of Theorem 3.7.

5.2. *More general regression functions:*

We have analyzed in this paper a stump based change point model: The model analyzed in this paper can be easily generalized to one where the levels (α_0, β_0) on either side of the boundary are replaced by some unknown functions of X. As an example, one may fit the following non-parametric model:

$$Y_i = f(X_i) \mathbb{1}_{X_i \le d_0} + g(X_i) \mathbb{1}_{X_i > d_0} + \xi_i$$

where both f, g are smooth and $f(d_0) \neq g(d_0)$. One may estimate f, g, d_0 using the following HEF:

$$(\hat{f}^k, \hat{g}^k, \hat{d}^k) = \underset{f,g,d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \tilde{H}_k \left(Y_i - f(X_i) \mathbb{1}_{X_i \le d} - g(X_i) \mathbb{1}_{X_i > d} \right)\,,$$

with (f, g) restricted to an appropriate class of functions (depending upon the underlying application). This model is well investigated in the literature using the squared error loss (e.g. see [28]), however the properties of the robust estimators (i.e. estimators obtained by minimizing HEF) are still largely unknown and worthy of investigation in the presence of heavy tailed errors.

5.3. Smoothed change plane problem:

The change plane estimators analyzed in Subsection 3.2 are NP-hard to compute as HEF is discontinuous at the change boundary. One may replace the indicator function involved in HEF by a smooth sigmoid function to estimate the unknown parameters as follows:

$$(\hat{\alpha}^k, \hat{\beta}^k, \hat{d}^k) = \underset{\alpha,\beta,d}{\operatorname{argmin}} \frac{1}{n} \sum_i \tilde{H}_k \left(Y_i - \alpha - (\beta - \alpha) \frac{e^{X_i^\top d_0/\sigma_n}}{1 + e^{X_i^\top d_0/\sigma_n}} \right)$$

for some bandwidth parameter $\sigma_n \to 0$ as $n \to \infty$. The sigmoid function converges to the indicator function as $n \to \infty$ and is differentiable with respect to (α, β, d), therefore one may employ gradient descent to estimate the parameters (e.g. similar to [19], [36] or [32]); however, as the loss function is non-convex, there is no guarantee that gradient descent type techniques initiated from a random point on the parameter surface will converge to a global minimum. One way to address this issue is to replace the indicator function by a convex surrogate (i.e. logit function as in logistic regression, exponential function as in adaboost), but as the convex function does not converge to the indicator function, it is unclear whether this method will lead to a consistent estimator of d_0. However, such methods merit deeper investigation as they may facilitate efficient computation of the change plane estimator.

Appendix A: Proofs of selected Theorems

For all proofs below and in the supplement we will assume $\alpha_0 > \beta_0$ for simplicity of presentation. The derivations all go through for the reverse inequality upon minor adjustments of the proofs presented in the paper.

A.1. Proof of Theorem 2.4

We divide the whole proofs into few supplementary lemmas, whose proofs can be found in Supplementary document. Our first lemma provides a lower bound on the the probability of a random walk staying always positive. We believe that this lemma has been proved before, but we were unable to find a proper source to cite. Hence we will provide our own proof in Supplementary document.

Lemma A.1. *Suppose* $\{S_i\}_{i\geq 0}$ *is a positively drifted random walk (i.e.* $S_i = \sum_{j=1}^{i}(X_i+\mu)$, $\mathbb{E}(X)=0, \mu>0$*) with* $S_0=0$*. Then we have:*

$$\mathbb{P}\left(\max_{1\leq i\leq n} S_i < 0\right) \geq \frac{1}{n}\mathbb{P}\left(S_n<0\right).$$

The compound Poisson process is essentially a two sided random walk where are number of steps till time t follows a Poisson process. Therefore, we start by establishing tail bound on the minimizer of the random walk and then relate it to the tail of minimizer of the compound Poisson process. Our next lemma establishes that if the step distribution of a random walk follows a Pareto distribution, then the minimizer of the random walk is also heavy-tailed:

Lemma A.2. *Suppose* $\xi_1, \xi_2, \ldots$ *i.i.d. random variables with the following distribution:*

$$\mathbb{P}\left(|\xi|>t\right) = \frac{1}{1+t^{\gamma}}$$

and $\mathbb{P}(\xi>t) = 1-\mathbb{P}(\xi\leq -t)$ *for all* $t>0$*. Define* $X_i=\xi_i+\mu$ *for some* $\mu>0$ *and a random walk based on* X_i*'s, i.e* $S_n=\sum_{i=1}^{n}X_i$*. Suppose* M *denotes the minimizer of the random walk on* $\mathbb{Z}^+$*. Then we have:*

$$\mathbb{P}\left(M\geq k\right) \geq \frac{c_1c_2p^*}{\gamma}\times\frac{1}{k^{\gamma}} := c_0k^{-\gamma},$$

for all $k\geq k_0 := 1\vee\lceil\mu^{-\gamma/(\gamma-1)}\rceil$*, where:*

1. $p^* = \mathbb{P}\left(S_i>0 \quad \forall \quad i\in\mathbb{N}\right) = \mathbb{P}(M=0)$.
2. $c_1 = \frac{1}{2(1+\mu^{-\gamma})(1+\mu)^{\gamma}}$.
3. $c_2 = \inf_{x\geq 1}\left(1-\frac{1}{1+x}\right)^{x-1}$.

The previous lemma indicates that the minimizer of the random walk with a heavy tailed step distribution is also heavy tailed. As the compound Poisson process is a two sided process (i.e. supported on entire real line), we next extend our lower bound on the tail of the minimizer of random walk obtained in previous lemma for a two sided random walk in the following lemma:

Lemma A.3. *Under the same structure as of Lemma A.2, we consider a two sided random walk with independent component on the either side. Define by* M_{ts} *as the minimizer of the two sided random walk and by* M_{os} *as the minimizer of one-sided random walk. Then we have:*

$$\mathbb{P}\left(|M_{ts}|\geq k\right) \geq 2p^*c_0k^{-\gamma}$$

for all $k\geq k_0$*, where* p^*, k_0, c_0 *are same as defined in Lemma A.2.*

Finally, we translate the lower bound on the tail of the minimizer of the two-sided random walk to the two sided compound Poisson process in the following lemma:

Lemma A.4. *Consider a two sided independent compound Poisson process with increment independent of the steps. More specifically, let* $\{X_i\}_{i\in\mathbb{N}}$ *be same as defined in Lemma A.2. Suppose* $\{X_i'\}_{i\in\mathbb{N}}$ *be an independent copy of* $\{X_i\}_{i\in\mathbb{N}}$. *Also suppose* $N_1(t)$ *and* $N_2(t)$ *are two independent Poisson process on* $\mathbb{R}^+$ *with some intensity function* $\Lambda(t)$. *The two sided independent compound Poisson process on* $\mathbb{R}$ *is defined as:*

$$X(t) = \begin{cases} \sum_{i=1}^{N_1(t)} X_i\,, & \text{if } t > 0 \\ \sum_{i=1}^{N_2(-t)} X_i'\,, & \text{if } t < 0 \\ 0\,, & \text{if } t = 0\,. \end{cases}$$

Let M *be the mid-argmin of* $X(t)$ *over* $\mathbb{R}$. *Then we have for all* $x > (k_0 + \gamma + \log 2)/f_X(d_0)$:

$$\mathbb{P}\left(M_{ts,CPP} > x\right) \geq \frac{c_0}{2f_X^{\gamma}(d_0)} x^{-\gamma}\,,$$

where c_0, k_0 *are same constants as defined in Lemma A.2.*

Combining Lemma A.2, A.3 and A.4, we conclude the proof of of lower bound on F_{ℓ_2}.

Now to prove the upper bound for F_{ℓ_1} we modify our arguments in the previous lemmas. Note that F_{ℓ_1} is the distribution of the minimizer of the following compound Poisson process:

$$CPP(t) = \sum_{i=1}^{N_+(t)} (\xi^* + \mu_0)\,\mathbb{1}_{t\geq 0} + \sum_{i=1}^{N_-(-t)} (\xi^* + \mu_0)\,\mathbb{1}_{t<0}$$

with $CPP(t) = 0$, N_+ and N_- are two independent Poisson processes as before and:

$$\xi^* \leftarrow \{|\xi + (\alpha_0 - \beta_0)| - |\xi|\} - \mathbb{E}\left[|\xi + (\alpha_0 - \beta_0)| - |\xi|\right]$$

with $\mu_0 = \mathbb{E}\left[|\xi + (\alpha_0 - \beta_0)| - |\xi|\right] > 0$. Before going into the details of the proof, we state Hoeffding's inequality bound:

$$\begin{aligned} \mathbb{P}\left(S_n < 0\right) &= \mathbb{P}\left(\sum_{i=1}^{n} \xi_i^* < -n\mu_0\right) \\ &= \mathbb{P}\left(\bar{\xi}_n^* < -\mu_0\right) \leq e^{-\frac{n\mu_0^2}{8(\alpha_0-\beta_0)^2}}\,. \end{aligned}$$

Here we will highlight the steps where a modification is needed. First note that, in case of one sided random walk (same situation as in Lemma A.2) we obtain the upper bound as follows:

$$\begin{aligned} P\left(M_{os} \geq k\right) &= \sum_{j\geq k} \mathbb{P}\left(M_{os} = j\right) \\ &= p^* \sum_{j\geq k} \mathbb{P}\left(\max_{1\leq i\leq j} S_i < 0\right) \end{aligned}$$

$$\begin{aligned}
&\leq p^* \sum_{j \geq k} \mathbb{P}\left(S_j < 0\right) \\
&\leq p^* \sum_{j \geq k} e^{-\frac{j\mu_0^2}{8(\alpha_0-\beta_0)^2}} \\
&= \frac{p^*}{1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}} e^{-\frac{k\mu_0^2}{8(\alpha_0-\beta_0)^2}} \qquad \text{(A.1)}
\end{aligned}$$

We now translate the tail bound on the minimizer of the one sided random walk to the minimizer of the two sided random walk as below:

$$\begin{aligned}
\mathbb{P}\left(M_{ts} = k\right) &= \mathbb{P}\left(S_K \leq S_i \ \forall\ 0 \leq i \leq k-1, S_k \leq S_i \ \forall\ k+1 \leq i < \infty, \right. \\
&\qquad \left. S_k \leq \inf_{j \geq 1} S_{-j}\right) \\
&\leq \mathbb{P}\left(S_K \leq S_i \ \forall\ 0 \leq i \leq k-1, S_k \leq S_i \ \forall\ k+1 \leq i < \infty\right) \\
&= \mathbb{P}\left(M_{os} = k\right) .
\end{aligned}$$

This, along with the upper bound on the tail of the one-sided random walk implies:

$$\mathbb{P}\left(M_{ts} \geq k\right) \leq \frac{p^*}{1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}} e^{-\frac{k\mu_0^2}{8(\alpha_0-\beta_0)^2}} .$$

Next, we translate the bound for the minimizer of a one-sided random walk to a one-sided compound Poisson process with steps $\xi^* + \mu_0$:

$$\begin{aligned}
P(M_{os,CPP} > x) &= \sum_{k=0}^{\infty} \mathbb{P}\left(M_{os,CPP} > x \mid N_1(x) = k\right) \mathbb{P}\left(N_1(x) = k\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}\left(\operatorname*{argmin}_{i \geq 0} S_i > k\right) \mathbb{P}\left(N_1(x) = k\right) \\
&\leq \frac{p^*}{1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}} \sum_{k=0}^{\infty} e^{-\frac{(k+1)\mu_0^2}{8(\alpha_0-\beta_0)^2}} \mathbb{P}\left(N_1(x) = k\right) \\
&= \frac{p^*}{1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}} e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} \sum_{k=0}^{\infty} e^{-\frac{k\mu_0^2}{8(\alpha_0-\beta_0)^2}} \frac{e^{-\Lambda(x)}\Lambda(x)^k}{k!} \\
&= \frac{p^*}{1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}} e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} e^{-\Lambda(x)} \sum_{k=0}^{\infty} \frac{\left(e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} \Lambda(x)\right)^k}{k!} \\
&= \frac{p^*}{e^{\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} - 1} \exp\left(-\Lambda(x)\left(1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}\right)\right) \\
&= \frac{p^*}{e^{\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} - 1} \exp\left(-x f_X\left(d_0\right)\left(1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}\right)\right),
\end{aligned}$$

and consequently for the two sided compound Poisson process:

$$\begin{aligned}\bar{F}_{\ell_1}(x) = \mathbb{P}\left(M_{ts,CPP} > x\right) &= \sum_{k=0}^{\infty} \mathbb{P}\left(M_{ts,CPP} > x \mid N_1(x) = k\right) \mathbb{P}(N_1(x) = k) \\ &= \sum_{k=0}^{\infty} \mathbb{P}\left(M_{ts} > k\right) \mathbb{P}(N_1(x) = k) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(M_{os} > k\right) \mathbb{P}(N_1(x) = k) \\ &= \mathbb{P}\left(M_{os,CPP} > x\right) \\ &\leq \frac{p^*}{e^{\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}} - 1} \exp\left(-x f_X\left(d_0\right)\left(1 - e^{-\frac{\mu_0^2}{8(\alpha_0-\beta_0)^2}}\right)\right).\end{aligned}$$

A.2. Proof of Theorem 3.1

Proof of lower bound: The following lemma, which is a finite sample analogue of the first conclusion of Theorem 2.4, is essential to establish the lower bound of Theorem 3.1:

Lemma A.5. *Suppose, for a fixed n, F_{n,ℓ_2} denotes the distribution of $n(\hat{d}_i^{\ell_2} - d_{0,i})$. Then we have for all $2\gamma/f_X(d_0) \leq |x| \leq (\delta_1 \wedge \delta_2)n$ (for some constants δ_1, δ_2 independent of n defined explicitly in the proof):*

$$1 - F_{n,\ell_2}(x) = \mathbb{P}\left(\left|n(\hat{d}_i^{\ell_2} - d_{0,i})\right| \geq x\right) \geq \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+2}} \times \left(\frac{f_{X,\max}}{1 - F_X(d_0)}\right)^{-\gamma} \times x^{-\gamma}$$

where $f_{X,\max}$ is the maximum value of the density of X and c_1, c_2, p^* are same as defined in Lemma A.2.

The proof of the Lemma can be found in Appendix B. For notational simplicity, set:

$$C = \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+2}} \times \left(\frac{f_{X,\max}}{1 - F_X(d_0)}\right)^{-\gamma}.$$

Using the above lemma we have:

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq i\leq m} \frac{n}{m^{1/\gamma}} \left|\hat{d}_i^{\ell_2} - d_{0,i}\right| > t\right) &= 1 - \mathbb{P}\left(\max_{1\leq i\leq m} \frac{n}{m^{1/\gamma}} \left|\hat{d}_i^{\ell_2} - d_{0,i}\right| \leq t\right) \\ &= 1 - \left(F_{n,\ell_2}(tm^{1/\gamma})\right)^m \\ &= 1 - \left(1 - \bar{F}_{n,\ell_2}(tm^{1/\gamma})\right)^m \\ &\geq 1 - \left(1 - C(tm^{1/\gamma})^{-\gamma}\right)^m\end{aligned}$$

$$= 1 - \left(1 - Cm^{-1}t^{-\gamma}\right)^m$$
$$\longrightarrow 1 - e^{-Ct^{-\gamma}}$$

Note that Lemma A.5 is applicable here as for any fixed t, $tm^{1/\gamma} \ll n$ because $m^{1/\gamma} \ll n$ and as $m \uparrow \infty$, $tm^{1/\gamma} \geq 2\gamma/f_X(d_0)$ for all large m. This completes the proof.

Proof of upper bound: The proof of upper bound relies on the following Lemma, which is a analogue of Lemma A.5, where we establish an upper bound on the finite sample distribution of $n(\hat{d}_i - d_{i,0})$ with bounded supported error distribution ξ:

Lemma A.6. *Let F_{n,ℓ_1} denotes the distribution of $\left|n(\hat{d}_i^{\ell_1} - d_{0,i})\right|$. Then we have for $0 \leq |x| \leq n\delta_1$ (for some constant δ_1 defined explicitly in the proof):*

$$1 - F_{n,\ell_1}(x) = \mathbb{P}\left(\left|n(\hat{d}_i^{\ell_1} - d_{0,i})\right| \geq x\right) \leq \frac{2e^{-c}}{1-e^{-c}} e^{-x\frac{f_X(d_0)}{2}(1-e^{-c})},$$

where $c = \mu^2/4b^2$, $\mu = \mathbb{E}[|\xi + (\alpha_0 - \beta_0)| - |\xi|]$ and b is the range of the random variable $(|\xi + (\alpha_0 - \beta_0)| - |\xi|) - \mu$.

Using the above lemma we have:

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq i\leq m} \frac{n}{\log m}\left|\hat{d}_i^{\ell_1} - d_{0,i}\right| > t\right) &\leq \sum_{i=1}^m \mathbb{P}\left(n\left|\hat{d}_i^{\ell_1} - d_{0,i}\right| > t\log m\right) \\
&\leq \frac{2e^{-c}}{1-e^{-c}} m e^{-t\log m\frac{f_X(d_0)}{2}(1-e^{-c})} \\
&\leq \frac{2e^{-c}}{1-e^{-c}} e^{-\log m\left(t\frac{f_X(d_0)}{2}(1-e^{-c})-1\right)}.
\end{aligned}$$

This completes the proof.

A.3. Proof of Theorem 3.3

To prove the lower bound, we consider a simple model: Assume that, for each problem, the true change point is 0 (i.e. $d_{i,0} = 0$ for all i), the covariates $X'_{i,j}s$ are all i.i.d. Uniform $(-1, 1)$ and error distribution is normal. We first observe that for any estimator $\hat{d}_i$ of $d_{i,0}$ we have:

$$\left|\hat{d}_i - d_{i,0}\right| \geq \min_{1\leq j\leq n} |X_{i,j} - d_{i,0}|,$$

i.e. we can't estimate a change point better than its closest order statistic. Note that when $d_{i,0} = 0$, we have:

$$\left|\hat{d}_i\right| \geq \min_{1\leq j\leq n} |X_{i,j}| = \min_{1\leq i\leq n} U_i$$

where $U_1, \ldots, U_n \overset{i.i.d.}{\sim} U(0,1)$. Hence to prove the theorem, all we need to show is that:

$$\liminf_{n,m\to\infty} \frac{1}{\log m}\mathbb{E}\left[\max_{1\le i\le m} nZ_{i:n}\right] \ge C > 0$$

where $Z_{i:n}$ are i.i.d with the common distribution being that of the minimum of n uniform $(0,1)$ random variables. Note that for any $0 \le t \le n$:

$$\begin{aligned}
\mathbb{P}\left(nZ_{i:n} \ge t\right) &= \mathbb{P}\left(\min_{1\le i\le n} U_i \ge \frac{t}{n}\right) \\
&= \left(1-\frac{t}{n}\right)^n \\
\implies \mathbb{P}\left(nZ_{i:n} \le t\right) &= 1-\left(1-\frac{t}{n}\right)^n := F_{nZ_{1:n}}(t)\,.
\end{aligned}$$

Therefore we have:

$$\begin{aligned}
\frac{1}{\log m}\mathbb{E}\left[\max_{1\le i\le m} nZ_{i:n}\right] &= \frac{1}{\log m}\int_0^n \mathbb{P}\left(\max_{1\le i\le m} nZ_{i:n} \ge t\right)\, dt \\
&= \frac{1}{\log m}\int_0^n \left[1-\mathbb{P}\left(\max_{1\le i\le m} nZ_{i:n} \le t\right)\right]\, dt \\
&= \frac{1}{\log m}\int_0^n \left[1-F^m_{nZ_{1:n}}(t)\right]\, dt \\
&= \frac{1}{\log m}\int_0^n \left[1-\left(1-\left(1-\frac{t}{n}\right)^n\right)^m\right]\, dt \qquad \text{(A.2)}
\end{aligned}$$

Next using the following inequality:

$$\left(1+\frac{t}{n}\right)^n \ge e^t\left(1-\frac{t^2}{n}\right) \quad \forall \quad |t| \le n\,,$$

we obtain from equation (A.2):

$$\begin{aligned}
\frac{1}{\log m}\mathbb{E}\left[\max_{1\le i\le m} nZ_{i:n}\right] &\ge \frac{1}{\log m}\int_0^n \left[1-\left(1-e^{-t}+e^{-t}\frac{t^2}{n}\right)^m\right]\, dt \\
&= \frac{1}{\log m}\int_0^n \left[1-\sum_{i=0}^m \binom{m}{i}\left(e^{-t}\frac{t^2}{n}\right)^i\left(1-e^{-t}\right)^{m-i}\right] \\
&= \frac{1}{\log m}\int_0^n \left[1-\left(1-e^{-t}\right)^m\right]\, dt \\
&\quad - \frac{1}{\log m}\int_0^n \sum_{i=1}^m \binom{m}{i}\left(e^{-t}\frac{t^2}{n}\right)^i\left(1-e^{-t}\right)^{m-i}\, dt \\
&= \frac{1}{\log m}\int_0^\infty \left[1-\left(1-e^{-t}\right)^m\right]\, dt \\
&\quad - \frac{1}{\log m}\int_n^\infty \left[1-\left(1-e^{-t}\right)^m\right]\, dt
\end{aligned}$$

$$- \frac{1}{\log m} \int_0^n \sum_{i=1}^{m} \binom{m}{i} \left(e^{-t} \frac{t^2}{n}\right)^i \left(1 - e^{-t}\right)^{m-i} \, dt$$
$$:= a_{m,1} - a_{m,2} - a_{m,3} \tag{A.3}$$

We next show that $a_{m,2}, a_{m,3} \longrightarrow 0$ and $a_{m,1} \longrightarrow 1$ as $m \to \infty$, as long as $n \geq 2m$. We start with $a_{m,3}$:

$$\begin{aligned}
a_{m,3} &= \frac{1}{\log m} \int_0^n \sum_{i=1}^{m} \binom{m}{i} \left(e^{-t} \frac{t^2}{n}\right)^i \left(1 - e^{-t}\right)^{m-i} \, dt \\
&\leq \frac{1}{\log m} \int_0^n \sum_{i=1}^{m} \binom{m}{i} \left(e^{-t} \frac{t^2}{n}\right)^i \, dt \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \int_0^n \left(e^{-t} \frac{t^2}{n}\right)^i \, dt \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \frac{1}{in^i} \int_0^n t^{2i} \, ie^{-it} \, dt \\
&\leq \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \frac{1}{in^i} \int_0^\infty t^{2i} \, ie^{-it} \, dt \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \frac{1}{in^i} \frac{(2i)!}{i^{2i}} \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \frac{(m-i+1)\cdots m}{in^i i^i} \frac{(i+1)\cdots(i+i)}{i^i} \\
&\leq \frac{1}{\log m} \sum_{i=1}^{m} \frac{(m-i+1)\cdots m}{i\left(\frac{n}{2}\right)^i i^i} \\
&\leq \frac{1}{\log m} \sum_{i=1}^{m} \left(\frac{2m}{n}\right)^i \frac{1}{i^{i+1}} \leq \frac{1}{\log m} \sum_{i=1}^{m} \frac{1}{i^{i+1}} \longrightarrow 0 \,.
\end{aligned}$$

For the other term $a_{n,2}$:

$$\begin{aligned}
|a_{n,2}| &= \left| \frac{1}{\log m} \int_n^\infty \left[1 - \left(1 - e^{-t}\right)^m\right] \, dt \right| \\
&= \left| \frac{1}{\log m} \int_n^\infty \left[1 - \sum_{i=0}^{m} \binom{m}{i} e^{-it} (-1)^i\right] \, dt \right| \\
&= \left| \frac{1}{\log m} \int_n^\infty \sum_{i=1}^{m} \binom{m}{i} e^{-it} (-1)^{i+1} \, dt \right| \\
&\leq \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \int_n^\infty e^{-it} \, dt
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \frac{e^{-ni}}{i} \\
&= \frac{2^m}{\log m} \mathbb{E}\left[\frac{e^{-nX}}{X} \mathbb{1}_{X \geq 1} \right] \qquad [X \sim \mathrm{Bin}(n,p)] \\
&\leq \frac{2^m}{\log m} \mathbb{E}\left[\frac{e^{-2mX}}{X} \mathbb{1}_{X \geq 1} \right] \qquad [\because n \geq 2m] \\
&\leq \frac{2^m}{\log m} \mathbb{E}\left[e^{-2mX} \right] \qquad [\because n \geq 2m] \\
&= \frac{2^m}{\log m} \left(\frac{1}{2} e^{-2m} + \frac{1}{2} \right)^m \\
&= \frac{1}{\log m} \left(e^{-2m} + 1 \right)^m \longrightarrow 0 \,.
\end{aligned}
$$

Now the calculation of $a_{n,1}$ is similar by replacing n with 0. We have:

$$
\begin{aligned}
a_{n,1} &= \frac{1}{\log m} \int_0^\infty \left[1 - \left(1 - e^{-t} \right)^m \right] \, dt \\
&= \frac{1}{\log m} \int_0^\infty \sum_{i=1}^{m} \binom{m}{i} e^{-it} (-1)^{i+1} \, dt \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \binom{m}{i} \frac{(-1)^{i+1}}{i} \\
&= \frac{1}{\log m} \sum_{i=1}^{m} \frac{1}{i} \overset{m \to \infty}{\longrightarrow} 1 \,.
\end{aligned}
$$

where the last equality follows from the representation of Harmonic number. Therefore from equation (A.3) we conclude:

$$
\liminf_{m,n \to \infty} \frac{1}{\log m} \mathbb{E}\left[\max_{1 \leq i \leq m} n Z_{i:n} \right] \geq 1 \,.
$$

This concludes the proof.

A.4. Proof of Theorem 3.6

A.4.1. Case 1: $0 \leq k < \infty$

We first establish the rate of convergence of $(\hat{\alpha}_{\text{init}}, \hat{\beta}_{\text{init}}, \hat{d})$. Towards that direction, we use the following semi-metric over the parameter space Θ:

$$
\mathsf{dist}(\theta_1, \theta_2) = \sqrt{(\alpha_1 - \alpha_2)^2 + (\beta_1 - \beta_2)^2 + \mathbb{P}\left(\mathrm{sign}(X^\top d) \neq \mathrm{sign}(X^\top d_0) \right)}
$$

The curvature of the population score function $\mathbb{M}(\theta)$ around the its value at minimizer $\mathbb{M}(\theta_0)$ is obtained via the similar calculation as in in the proof of

Theorem 2.1 (specifically equation (A.16) in the supplement). Consider all $\theta \in \Theta$ such that $\mathsf{dist}(\theta, \theta_0) \leq \delta$ where δ is such that $|\alpha_0 - \beta_0| > 2\delta$. For such that θ we have:

$$\begin{aligned}
\mathbb{M}(\theta) - \mathbb{M}(\theta_0) &= \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X^\top d \vee X^\top d_0 \leq 0\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \alpha_0 - \beta\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X^\top d_0 < 0 < X^\top d\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \beta_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X^\top d < 0 < X^\top d_0\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \beta_0 - \beta\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X^\top d \wedge X^\top d_0 > 0\right) \\
&\geq \frac{C_k}{2}\left[(\alpha_0 - \alpha)^2 \mathbb{P}\left(X^\top d \vee X^\top d_0 \leq 0\right)\right) \\
&\qquad + (\beta_0 - \beta)^2 \mathbb{P}\left(X^\top d \wedge X^\top d_0 > 0\right) \\
&\qquad + \mathbb{P}\left(\operatorname{sign}(X^\top d) \neq \operatorname{sign}(X^\top d_0)\right)\left\{2(\alpha_0 - \beta_0 - \delta)^2\right\}\Big] \\
&\geq C_k\left[(\alpha_0 - \alpha)^2 + (\beta_0 - \beta)^2 + \mathbb{P}\left(\operatorname{sign}(X^\top d) \neq \operatorname{sign}(X^\top d_0)\right)\right] \\
&= C_k \mathsf{dist}^2(\theta, \theta_0)\,. \qquad \text{(A.4)}
\end{aligned}$$

Consistency: We use argmin continuous mapping theorem (Theorem 3.2.2 of [38]) to establish the consistency of the initial estimator. As the parameter space is bounded, our estimates are by default tight. As the process $\mathbb{M}(\theta) - \mathbb{M}(\theta_0)$ is continuous with respect to θ and has a clear minima at $\theta = \theta_0$ all we need to show for any compact subset $K \subseteq \Theta$:

$$\sup_{\theta \in K} |(\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) - (\mathbb{M}(\theta) - \mathbb{M}(\theta_0))| = o_p(1)\,.$$

Consider a collection of functions $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where the individual functions $f_\theta(X, \xi)$ is defined as:

$$\begin{aligned}
f_\theta(X, \xi) &= \left(\tilde{H}_k\left(\xi + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \vee X^\top d_0 \leq 0} \\
&\qquad \left(\tilde{H}_k\left(\xi + \alpha_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d_0 \leq 0 < X^\top d} \\
&\qquad + \left(\tilde{H}_k\left(\xi + \beta_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \leq 0 < X^\top d_0} \\
&\qquad + \left(\tilde{H}_k\left(\xi + \beta_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \wedge X^\top d_0 > 0} \\
&:= \sum_{i=1}^{4} g_\theta^i(\xi) h_\theta^i(X)
\end{aligned}$$

with:

$$\begin{aligned}
g_\theta^1(\xi) &= \left(\tilde{H}_k\left(\xi + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi)\right), \qquad & h_\theta^1(X) &= \mathbb{1}_{X^\top d \vee X^\top d_0 \leq 0}\,, \\
g_\theta^2(\xi) &= \left(H_k\left(\xi + \alpha_0 - \beta\right) - \tilde{H}_k(\xi)\right), \qquad & h_\theta^2(X) &= \mathbb{1}_{X^\top d_0 \leq 0 < X^\top d}\,,
\end{aligned}$$

$$g^3_\theta(\xi) = \left(\tilde{H}_k\left(\xi + \beta_0 - \alpha\right) - \tilde{H}_k(\xi)\right), \qquad h^3_\theta(X) = \mathbb{1}_{X^\top d \le 0 < X^\top d_0},$$
$$g^4_\theta(\xi) = \left(\tilde{H}_k\left(\xi + \beta_0 - \beta\right) - \tilde{H}_k(\xi)\right), \qquad h^4_\theta(X) = \mathbb{1}_{X^\top d \wedge X^\top d_0 > 0}.$$

As the Huber function H_k is Lipschitz with Lipschitz constant k, our criterion function $\tilde{H}_k$ is Lipschitz with Lipschitz constant $(k+1)$. As our parameter space is compact, the functions $\{g^i_\theta, h^i_\theta\}_{1\le i\le 4}$ are uniformly bounded, and has constant envelope, say F. That the functions $\{g^i_\theta\}_{\theta\in\Theta}$ for $i = 1, 2, 3, 4$ has finite VC dimension v (i.e. does not grow with n or p) is immediate. On the other hands, as all the p-dimensional hyperplanes passing through origin has VC dimension p. Hence the functions $\{h^i_\theta\}_{\theta\in\Theta}$ has VC dimension p. Define $\mathcal{F}_{g,i} = \{g^i_\theta : \theta \in \Theta\}$ for $1 \le i \le 4$ and $\mathcal{F}_{h,i} = \{h^i_\theta : \theta \in \Theta\}$ for $1 \le i \le 4$. Combining these we obtain:

$$\begin{aligned}
\sup_Q N(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)) &\le \sup_Q N\left(\epsilon\|F\|_{Q,1}, \sum_{i=1}^4 \mathcal{F}_{g,i}\mathcal{F}_{h,i}, L_1(Q)\right) \\
&\le \Pi_{i=1}^4 \sup_Q N(\epsilon\|F\|_{Q,1}, \mathcal{F}_{g,i}\mathcal{F}_{h,i}, L_1(Q)) \\
&\le \Pi_{i=1}^4 KVC(\mathcal{F}_{g,i})VC(\mathcal{F}_{h,i})(16e)^{VC(\mathcal{F}_{g,i})+VC(\mathcal{F}_{h,i})} \\
&\qquad \times \left(\frac{1}{\epsilon}\right)^{(VC(\mathcal{F}_{g,i})+VC(\mathcal{F}_{h,i})-2)} \\
&\le \Pi_{i=1}^4 Kvp(16e)^{v+p}\left(\frac{1}{\epsilon}\right)^{(v+p-2)} \\
&= K^4(16e)^{4(v+p)}\left(\frac{1}{\epsilon}\right)^{4(v+p-2)} \\
&= K^4\left(\frac{16e}{\epsilon}\right)^{4(v+p)}
\end{aligned}$$

This along with the fact $p/n \to 0$ implies:

$$\frac{1}{n}\log\left(\sup_Q N(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)))\right) \to 0.$$

Therefore $\mathcal{F}$ is Glivenko-Cantelli class of functions and using Theorem 2.4.3 of [38] we conclude that:

$$\sup_{\theta\in K} |(\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) - (\mathbb{M}(\theta) - \mathbb{M}(\theta_0))| = \|(\mathbb{P}_n - P)\|_{\mathcal{F}} = o_p(1).$$

This establishes the consistency of $(\hat{\alpha}_{\text{init}}, \hat{\beta}_{\text{init}}, \hat{d})$.

Rate of convergence of initial estimators: So far we have established the quadratic curvature of $\mathbb{M}(\theta)$ around its unique minimizer θ_0 and also the consistency of $\hat{\theta}_{\text{init}} = (\hat{\alpha}_{\text{init}}, \hat{\beta}_{\text{init}}, \hat{d})$. In this section we show that:

$$\sqrt{\frac{n}{p}}\left(\log\frac{n}{p}\right)^{-\frac{1}{2}} \mathsf{dist}(\hat{\theta}_{\text{init}}, \theta_0) = O_p(1)$$

which (along with Assumption 3.5) implies:

$$\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\left[(\hat{\alpha}_{\text{init}}-\alpha_0)^2+\left(\hat{\beta}_{\text{init}}-\beta_0\right)^2+\left\|\hat{d}-d_0\right\|\right]=O_p(1)\,.$$

To establish the rate, all is left to do is to find a bound on the modulus of continuity of the empirical process $\mathbb{M}_n(\theta)$, i.e we need to find $\phi_n(\delta)$ such that:

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0)\le\delta}|(\mathbb{M}_n-\mathbb{M})(\theta_0)-(\mathbb{M}_n-\mathbb{M})(\theta)|\right]\le\frac{\phi_n(\delta)}{\sqrt{n}}\,. \tag{A.5}$$

Towards that end, define a local collection of functions $\mathcal{F}_\delta=\{f_\theta : d(\theta,\theta_0)\le\delta\}$. Note that when $d(\theta,\theta_0)\le\delta$ we have:

$$\max\left\{|\alpha-\alpha_0|,|\beta-\beta_0|,\sqrt{\mathbb{P}\left(\text{sign}(X^\top d)\neq\text{sign}(X^\top d_0)\right)}\right\}\le\delta\,.$$

For any such θ we have:

$$\begin{aligned}
\mathbb{E}\left[f_\theta(X,\xi)^2\right]&=\mathbb{E}\left(\tilde{H}_k\left(\xi+\alpha_0-\alpha\right)-\tilde{H}_k(\xi)\right)^2\mathbb{P}\left(X^\top d\vee X^\top d_0\le 0\right)\\
&\quad\mathbb{E}\left(\tilde{H}_k\left(\xi+\alpha_0-\beta\right)-\tilde{H}_k(\xi)\right)^2\mathbb{P}\left(X^\top d_0\le 0<X^\top d\right)\\
&\quad+\mathbb{E}\left(\tilde{H}_k\left(\xi+\beta_0-\alpha\right)-\tilde{H}_k(\xi)\right)^2\mathbb{P}\left(X^\top d\le 0<X^\top d_0\right)\\
&\quad+\mathbb{E}\left(\tilde{H}_k\left(\xi+\beta_0-\beta\right)-\tilde{H}_k(\xi)\right)^2\mathbb{P}\left(X^\top d\wedge X^\top d_0>0\right)\\
&\lesssim C_k\left[\delta^2+\mathbb{P}\left(\text{sign}(X^\top d)\neq\text{sign}(X^\top d_0)\right)\right]\\
&\lesssim C_k\delta^2
\end{aligned} \tag{A.6}$$

Hence applying Theorem 8.7 of [35] we conclude:

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0)\le\delta}|(\mathbb{M}_n-\mathbb{M})(\theta_0)-(\mathbb{M}_n-\mathbb{M})(\theta)|\right]\lesssim\sqrt{\frac{p}{n}}\delta\sqrt{\log\frac{1}{\delta}}\vee\frac{p}{n}\log\frac{1}{\delta}$$

Therefore a valid choice of ϕ_n in equation (A.5) is:

$$\phi_n(\delta)=\sqrt{p}\delta\sqrt{\log\frac{1}{\delta}}\vee\frac{p}{\sqrt{n}}\log\frac{1}{\delta}\,.$$

Using this ϕ_n in Theorem 3.4.1 of [38] we conclude that:

$$\sqrt{\frac{n}{p}}\left(\log\frac{n}{p}\right)^{-\frac{1}{2}}d(\hat{\theta}_{\text{init}},\theta_0)=O_p(1)\,.$$

Finally, as the function class under consideration $\mathcal{F}=\{f_\theta:\theta\in\Omega\times S^{p-1}\}$ is uniformly bounded, an application of Theorem 2 of [30] yields:

$$\mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}d(\hat{\theta}_{\text{init}},\theta_0)\ge t\right)\le C_k e^{-c_k t} \tag{A.7}$$

for some constants $C_k, c_k > 0$ which depends on k. This in particular implies that:

$$\mathbb{E}\left[\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\mathbb{P}\left(\text{sign}(X^\top\hat{d}) \neq \text{sign}(X^\top d_0)\right)\right] \leq C_k \tag{A.8}$$

for some constant $C_k > 0$ depends on k.

Rate of convergence of the final estimators: We now present the proof that the rate of convergence of the final estimator. The proof for $\hat{\alpha}$ and $\hat{\beta}$ are similar and therefore we only the present the proof for $\hat{\alpha}$. Before delving into the technical details, we introduce some notation:

$$\begin{aligned}
\mathbb{M}_n(\alpha, d) &= \frac{1}{n}\sum_{i=1}^{n}\tilde{H}_k(Y_i - \alpha)\mathbb{1}_{X_i^\top d \leq 0} \\
\mathbb{R}_n(\alpha, d_1, d_2) &= \frac{1}{n}\sum_{i=1}^{n}\tilde{H}_k(Y_i - \alpha)\left(\mathbb{1}_{X_i^\top d_1 \leq 0} - \mathbb{1}_{X_i^\top d_2 \leq 0}\right) \\
\mathbb{M}(\alpha, d) &= \mathbb{E}[\mathbb{M}_n(\alpha, d)] = \mathbb{E}\left[\tilde{H}_k(Y - \alpha)\mathbb{1}_{X^\top d \leq 0}\right].
\end{aligned}$$

From Lemma A.8 we have for all $|\alpha - \alpha_0| \leq \eta$ (for some small enough $\eta > 0$):

$$\mathbb{M}(\alpha, d_0) - \mathbb{M}(\alpha_0, d_0) \geq C_k(\alpha - \alpha_0)^2$$

In terms of the processes introduced above, we can write our final estimator $\hat{\alpha}$ as:

$$\begin{aligned}
\hat{\alpha} &= \underset{\alpha}{\text{argmin}}\,\frac{1}{n}\sum_{i=1}^{n}\tilde{H}_k(Y_i - \alpha)\mathbb{1}_{X_i^\top \hat{d} \leq 0} \\
&= \underset{\alpha}{\text{argmin}}\,\mathbb{M}_n(\alpha, \hat{d}) \\
&= \underset{\alpha}{\text{argmin}}\left[\mathbb{M}_n(\alpha, d_0) + \mathbb{R}_n(\alpha, \hat{d}, d_0)\right]
\end{aligned} \tag{A.9}$$

Consistency of the above estimator follows from the similar calculation as of its previous incarnation, hence we skip it here for brevity. The remainder term $\mathbb{R}_n$ can be bounded as:

$$\begin{aligned}
\sup_{\alpha:|\alpha-\alpha_0|\leq\delta}\left|\mathbb{R}_n(\alpha, \hat{d}, d_0) - \mathbb{R}_n(\alpha_0, \hat{d}, d_0)\right| &\leq \frac{k\delta}{n}\sum_{i=1}^{n}\mathbb{1}_{\text{sign}(X_i^\top\hat{d}) \neq \text{sign}(X_i^\top d_0)} \\
&= k\delta\mathbb{P}_n f_{\hat{d}} \\
&= k\delta\left[(\mathbb{P}_n - P)f_{\hat{d}} + Pf_{\hat{d}}\right]
\end{aligned}$$

Fix $\epsilon > 0$. Previously, we have established that:

$$Pf_{\hat{d}} = O_p\left(\frac{p}{n}\log\frac{n}{p}\right).$$

Therefore we can find t_0 such that for all $t > t_0$:

$$\mathbb{P}\left(Pf_{\hat{d}} > t_0 \frac{p}{n}\log\frac{n}{p}\right) \leq \epsilon\,.$$

Next we bound the fluctuation of the centered empirical process.

$$\begin{aligned}
&\mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}(\mathbb{P}_n - P)f_{\hat{d}} > t\right)\\
&\leq \mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}(\mathbb{P}_n - P)f_{\hat{d}} > t, Pf_{\hat{d}} \leq t_0\frac{p}{n}\log\frac{n}{p}\right) + \mathbb{P}\left(Pf_{\hat{d}} > t_0\frac{p}{n}\log\frac{n}{p}\right)\\
&\leq \mathbb{P}\left(\sup_{d: Pf_d \leq t_0 \frac{p}{n}\log\frac{n}{p}} \frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\|(\mathbb{P}_n - P)f_d\| > t\right) + \epsilon\\
&\leq \frac{n}{p}\left(\log\frac{n}{p}\right)^{-1} \times \frac{1}{t}\mathbb{E}\left[\sup_{d: Pf_d \leq t_0 \frac{p}{n}\log\frac{n}{p}} \|(\mathbb{P}_n - P)f_d\|\right] + \epsilon\\
&\leq \frac{n}{p}\left(\log\frac{n}{p}\right)^{-1} \times \frac{\sqrt{t_0}\frac{p}{n}\sqrt{\log\frac{n}{p}\log\frac{n}{t_0 p\log(n/p)}} \vee \frac{p}{n}\log\frac{n}{t_0 p\log(n/p)}}{t} + \epsilon\\
&= \frac{\sqrt{t_0}}{t}\sqrt{\frac{\log(n/p) - \log(t_0\log(n/p))}{\log(n/p)}} \vee \frac{1}{t}\frac{\log(n/p) - \log(t_0\log(n/p))}{\log(n/p)} + \epsilon
\end{aligned}$$

where the second last inequality follows from Theorem 8.7 of [35]. Therefore we have:

$$\limsup_{n\to\infty}\mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}(\mathbb{P}_n - P)f_{\hat{d}} > t\right) \leq \frac{\sqrt{t_0}}{t} \vee \frac{1}{t} + \epsilon\,,$$

which implies that for any fixed $\epsilon > 0$, we can t large enough to ensure that:

$$\limsup_{n\to\infty}\mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}(\mathbb{P}_n - P)f_{\hat{d}} > t\right) \leq 2\epsilon$$

Hence we have:

$$\sup_{|\alpha-\alpha_0|\leq\delta}\left|\mathbb{R}_n(\alpha,\hat{d},d_0) - \mathbb{R}_n(\alpha_0,\hat{d},d_0)\right| = \delta \times O_p\left(\frac{p}{n}\log\frac{n}{p}\right)\,.$$

Again fix $\epsilon > 0$ and choose $t_0 > 0$ such that:

$$\limsup_{n\to\infty}\mathbb{P}\left(\frac{n}{p}\left(\log\frac{n}{p}\right)^{-1}\sup_{|\alpha-\alpha_0|\leq\delta}\left|\mathbb{R}_n(\alpha,\hat{d},d_0) - \mathbb{R}_n(\alpha_0,\hat{d},d_0)\right| \geq \delta t_0\right) \leq \epsilon\,.$$

Also note that this t_0 only depends on ϵ, not δ. Henceforth define $r_n = \sqrt{n} \wedge (n/p)(\log(n/p))^{-1}$, the desired rate of convergence for the second stage estimator for notational simplicity. Further, by an application of Lemma 2.14.1 of [38]

we have for any $\delta > 0$:

$$\mathbb{E}\left[\sup_{|\alpha-\alpha_0|\le\delta} |(\mathbb{M}_n - \mathbb{M})(\alpha, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0)|\right] \lesssim \frac{\delta}{\sqrt{n}}\,. \tag{A.10}$$

Using a shelling type of argument, we have for any $t > 0$:

$$\begin{aligned}
&\mathbb{P}\left(r_n|\hat\alpha - \alpha_0| > t\right)\\
&\le \mathbb{P}\left(r_n|\hat\alpha - \alpha_0| > t, |\hat\alpha - \alpha_0| \le \eta\right) + \mathbb{P}\left(|\hat\alpha - \alpha_0| > \eta\right)\\
&\le \mathbb{P}\left(r_n|\hat\alpha - \alpha_0| > t, |\hat\alpha - \alpha_0| \le \eta\right) + \epsilon\\
&\le \mathbb{P}\left(\sup_{\alpha: tr_n^{-1} < |\alpha-\alpha_0| \le \eta} \Big\{\mathbb{M}_n(\alpha_0, d_0) + \mathbb{R}_n(\alpha_0, \hat d, d_0)\right.\\
&\qquad\qquad \left.-\mathbb{M}_n(\alpha, d_0) - \mathbb{R}_n(\alpha, \hat d, d_0)\Big\} \ge 0\right) + \epsilon\\
&\le \mathbb{P}\left(\sup_{\alpha: tr_n^{-1} < |\alpha-\alpha_0| \le \eta} \Big\{\mathbb{M}_n(\alpha_0, d_0) + \mathbb{R}_n(\alpha_0, \hat d, d_0)\right.\\
&\qquad \left.-\mathbb{M}_n(\alpha, d_0) - \mathbb{R}_n(\alpha, \hat d, d_0)\Big\} \ge 0, \mathbb{P}_n f_{\hat d} \le t_0 \frac{p}{n}\log\frac{n}{p}\right)\\
&\qquad + \mathbb{P}\left(\mathbb{P}_n f_{\hat d} > t_0 \frac{p}{n}\log\frac{n}{p}\right) + \epsilon\\
&\le \mathbb{P}\left(\sup_{\alpha: tr_n^{-1} < |\alpha-\alpha_0| \le \eta} \Big\{\mathbb{M}_n(\alpha_0, d_0) + \mathbb{R}_n(\alpha_0, \hat d, d_0)\right.\\
&\qquad \left.-\mathbb{M}_n(\alpha, d_0) - \mathbb{R}_n(\alpha, \hat d, d_0)\Big\} \ge 0, \mathbb{P}_n f_{\hat d} \le t_0 \frac{p}{n}\log\frac{n}{p}\right) + 2\epsilon\\
&\le \sum_{j=1}^{\log_2(\eta r_n/t)} \mathbb{P}\left(\sup_{\alpha: 2^{j-1}tr_n^{-1} < |\alpha-\alpha_0| \le 2^j tr_n^{-1}} \Big\{\mathbb{M}_n(\alpha_0, d_0) + \mathbb{R}_n(\alpha_0, \hat d, d_0)\right.\\
&\qquad\qquad \left.-\mathbb{M}_n(\alpha, d_0) - \mathbb{R}_n(\alpha, \hat d, d_0)\Big\} \ge 0, \mathbb{P}_n f_{\hat d} \le t_0 \frac{p}{n}\log\frac{n}{p}\right) + 2\epsilon\\
&\le \sum_{j=1}^{\log_2(\eta r_n/t)} \mathbb{P}\left(\sup_{\alpha: 2^{j-1}tr_n^{-1} < |\alpha-\alpha_0| \le 2^j tr_n^{-1}} \{\mathbb{M}_n(\alpha_0, d_0) - \mathbb{M}_n(\alpha, d_0)\}\right.\\
&\left. + \sup_{2^{j-1}tr_n^{-1} < |\alpha-\alpha_0| \le 2^j tr_n^{-1}} \left|\mathbb{R}_n(\alpha, \hat d, d_0) - \mathbb{R}_n(\alpha_0, \hat d, d_0)\right| \ge 0, \mathbb{P}_n f_{\hat d} \le t_0 \frac{p}{n}\log\frac{n}{p}\right) + 2\epsilon\\
&\le \sum_{j=1}^{\log_2(\eta r_n/t)} \mathbb{P}\left(\sup_{\alpha: 2^{j-1}tr_n^{-1} < |\alpha-\alpha_0| \le 2^j tr_n^{-1}} \{\mathbb{M}_n(\alpha_0, d_0) - \mathbb{M}_n(\alpha, d_0)\}\right.\\
&\qquad\qquad \left. + 2^j tr_n^{-1} t_0 \frac{p}{n}\log\frac{n}{p} \ge 0\right) + 2\epsilon\\
&\le \sum_{j=1}^{\log_2(\eta r_n/t)} \mathbb{P}\left(\sup_{\alpha: 2^{j-1}tr_n^{-1} < |\alpha-\alpha_0| \le 2^j tr_n^{-1}} \{\mathbb{M}_n(\alpha_0, d_0) - \mathbb{M}_n(\alpha, d_0)\}\right.\\
&\qquad\qquad \left. + 2^j tr_n^{-1} t_0 \frac{p}{n}\log\frac{n}{p} \ge 0\right) + 2\epsilon
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \mathbb{P}\Bigg(\sup_{\alpha: 2^{j-1} t r_n^{-1} < |\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\} \\
&\qquad + 2^j t r_n^{-1} t_0 \frac{p}{n} \log \frac{n}{p} \geq \inf_{\alpha: 2^{j-1} t r_n^{-1} < |\alpha-\alpha_0| \leq 2^j t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0)) \Bigg) + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \left[\frac{\mathbb{E}\left[\sup_{2^{j-1} t r_n^{-1} < |\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\}\right]}{\inf_{\alpha: 2^{j-1} t r_n^{-1} < |\alpha-\alpha_0| \leq 2^j t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right. \\
&\qquad \left. + \frac{2^j t r_n^{-1} t_0 \frac{p}{n} \log \frac{n}{p}}{\inf_{\alpha: 2^{j-1} t r_n^{-1} < |\alpha-\alpha_0| \leq 2^j t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right] + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \left[\frac{\mathbb{E}\left[\sup_{|\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\}\right]}{\inf_{\alpha: |\alpha-\alpha_0| \geq 2^{j-1} t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right. \\
&\qquad \left. + \frac{2^j t r_n^{-1} t_0 \frac{p}{n} \log \frac{n}{p}}{\inf_{\alpha: |\alpha-\alpha_0| \geq 2^{j-1} t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right] + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \left[\frac{\mathbb{E}\left[\sup_{|\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\}\right]}{\inf_{\alpha: |\alpha-\alpha_0| \geq 2^{j-1} t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right. \\
&\qquad \left. + \frac{2^j t t_0 r_n^{-2}}{\inf_{\alpha: |\alpha-\alpha_0| \geq 2^{j-1} t r_n^{-1}} (\mathbb{M}(\alpha) - \mathbb{M}(\alpha_0))} \right] + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \left[\frac{\mathbb{E}\left[\sup_{|\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\}\right]}{2^{2(j-1)} t^2 r_n^{-2}} \right. \\
&\qquad \left. + \frac{2^j t t_0 r_n^{-2}}{2^{2(j-1)} t^2 r_n^{-2}} \right] + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \frac{\mathbb{E}\left[\sup_{|\alpha-\alpha_0| \leq 2^j t r_n^{-1}} \{(\mathbb{M}_n - \mathbb{M})(\alpha_0, d_0) - (\mathbb{M}_n - \mathbb{M})(\alpha, d_0)\}\right]}{2^{2(j-1)} t^2 r_n^{-2}} + \frac{t_0}{t} + 2\epsilon \\
&\leq \sum_{j=1}^{\log_2(\eta r_n/t)} \frac{2^j t r_n^{-1}}{\sqrt{n} 2^{2(j-1)} t^2 r_n^{-2}} + \frac{t_0}{t} + 2\epsilon \qquad \text{[From equation (A.10)]} \\
&\leq \frac{1}{t} + \frac{t_0}{t} + 2\epsilon\,.
\end{aligned}
$$

Taking t large enough we conclude:

$$\limsup_{n\to\infty} \mathbb{P}\left(r_n|\hat{\alpha} - \alpha_0| > t\right) \leq 3\epsilon$$

This completes the proof.

A.4.2. Case 2: $k = \infty$, i.e. squared error loss

The proof for $k = \infty$ is similar to that of $0 \leq k < \infty$, the only difference is that the collection of functions $\mathcal{F}$ defined in the proof of the previous part is no

longer bounded. Hence we need to modify some parts of the proof carefully to take care of that.

Consistency: Consider the same function class $\mathcal{F}$ as in paragraph A.4.1. Note that now any individual function f_θ is:

$$\begin{aligned} f_\theta(X,\xi) = &\left(\xi(\alpha_0-\alpha)+\frac{1}{2}(\alpha_0-\alpha)^2\right)\mathbb{1}_{X^\top d\vee X^\top d_0\le 0} \\ &\left(\xi(\alpha_0-\beta)+\frac{1}{2}(\alpha_0-\beta)^2\right)\mathbb{1}_{X^\top d_0\le 0<X^\top d} \\ &+\left(\xi(\beta_0-\alpha)+\frac{1}{2}(\beta_0-\alpha)^2\right)\mathbb{1}_{X^\top d\le 0<X^\top d_0} \\ &+\left(\xi(\beta_0-\beta)+\frac{1}{2}(\beta_0-\beta)^2\right)\mathbb{1}_{X^\top d\wedge X^\top d_0>0} \end{aligned}$$

The envelope function F of $\mathcal{F}$ is as follows:

$$\begin{aligned} \sup_{\theta\in\Theta}|f_\theta(X,\xi)| &\le \sup_{(\alpha,\beta)\in\Omega}\Big[|\xi|\max\left\{|\alpha-\alpha_0|,|\alpha-\beta_0|,|\beta-\alpha_0|,|\beta-\beta_0|\right\} \\ &\qquad +\frac{1}{2}\left(\max\left\{|\alpha-\alpha_0|,|\alpha-\beta_0|,|\beta-\alpha_0|,|\beta-\beta_0|\right\}\right)^2\Big] \\ &\le C|\xi|+\frac{C^2}{2}:=F(X,\xi) \end{aligned}$$

The envelope function is integrable and following same analysis as of paragraph A.4.1 we conclude:

$$\sup_Q N\left(\epsilon\|F\|_{Q,1},\mathcal{F},L_1(Q)\right)\le K^4\left(\frac{16e}{\epsilon}\right)^{v+p}.$$

Hence $\mathcal{F}$ is a Glivenko-Cantelli class of functions and consistency follows from Theorem 2.4.3 of [38].

Rate of convergence of the initial estimate: To control the modulus of continuity, we can no longer apply Theorem 8.7 of [35] directly here as the functions are not uniformly bounded. Here we use the following modified version of Theorem 1 of [17]:

Proposition A.7. *Suppose* $\{\xi_1,\dots,\xi_n\}$ *are independent of random variables of* $\{X_1,\dots,X_n\}$ *and moreover* $\{X_1,\dots,X_n\}$ *are permutation invariant. Assume further that there exists a non-decreasing concave function* $\varphi_n:\mathbb{R}_+\to\mathbb{R}_+$ *with* $\varphi_n(0)=0$ *and constant* $b_n>0$ *such that for* $1\le k\le n$*:*

$$\mathbb{E}\left\|\sum_{i=1}^k\epsilon_i f(X_i)\right\|_{\mathcal{F}}\le\varphi_n(k)+b_n$$

for some i.i.d Rademacher random variables $\epsilon_1, \ldots, \epsilon_n$. *Then we have:*

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi_i f(X_i)\right\|_{\mathcal{F}} \leq 4\int_0^\infty \varphi_n\left(\sum_{i=1}^{n}\mathbb{P}\left(|\xi_i|>t\right)\right)\,dt + 2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right].$$

The proof of this proposition can be found in the Section B of the supplementary document. To apply the above proposition, we define

$$f_\theta(X_i,\xi_i) = \xi_i f_{\theta,1}(X_i) + f_{\theta,2}(X_i)$$

where:

$$\begin{aligned}
f_{\theta,1}(X_i) &= (\alpha_0-\alpha)\mathbb{1}_{X^\top d\vee X^\top d_0\leq 0} + (\alpha_0-\beta)\mathbb{1}_{X^\top d_0\leq 0<X^\top d} \\
&\quad + (\beta_0-\alpha)\mathbb{1}_{X^\top d\leq 0<X^\top d_0} + (\beta_0-\beta)\mathbb{1}_{X^\top d\wedge X^\top d_0>0} \\
f_{\theta_2}(X_i) &= \frac{1}{2}(\alpha_0-\alpha)^2\mathbb{1}_{X^\top d\vee X^\top d_0\leq 0} + \frac{1}{2}(\alpha_0-\beta)^2\mathbb{1}_{X^\top d_0\leq 0<X^\top d} \\
&\quad + \frac{1}{2}(\beta_0-\alpha)^2\mathbb{1}_{X^\top d\leq 0<X^\top d_0} + \frac{1}{2}(\beta_0-\beta)^2\mathbb{1}_{X^\top d\wedge X^\top d_0>0}
\end{aligned}$$

Both the collections $\mathcal{F}_1 = \{f_{\theta,1} : d(\theta,\theta_0)\leq\delta\}$ and $\mathcal{F}_2 = \{f_{\theta,2} : d(\theta,\theta_0)\leq\delta\}$ are uniformly bounded with VC dimension of the order p. It is also immediate that $Pf_{\theta,j}^2 \lesssim \delta^2$ for all $\theta : d(\theta,\theta_0)\leq\delta$, for $j\in\{1,2\}$. Hence we have from Theorem 8.7 of [35] for any $1\leq k\leq n$ and $\epsilon_1,\ldots.\epsilon_n$ i.i.d Rademacher random variables:

$$\begin{aligned}
\mathbb{E}\left[\sup_{d(\theta,\theta_0)\leq\delta}\left|\sum_{i=1}^{k}\epsilon_i f_{\theta,j}(X_i)\right|\right] &\leq L\left(\delta\sqrt{k}\sqrt{p\log\frac{AU}{\delta}} + pU\log\frac{AU}{\delta}\right) \qquad \text{(A.11)}\\
&:= \varphi_n(k) + b_n
\end{aligned}$$

for some constants $L>0, A>e^2$ and U is the uniform bounds on the individual functions and $\varphi_n(k) = L\delta\sqrt{k}\sqrt{p\log(AU/\delta)}$ and $b_n = pU\log(AU/\delta)$. Therefore using Proposition A.7:

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i=1}^{k}\xi_i f(X_i)\right\|_{\mathcal{F}_1} \qquad \text{(A.12)}\\
&\leq 4\int_0^\infty \varphi_n\left(n\mathbb{P}\left(|\xi_1|>t\right)\right)\,dt + 2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] \\
&\leq 4L\delta\sqrt{n}\sqrt{p\log\left(\frac{AU}{\delta}\right)}\int_0^\infty\sqrt{\mathbb{P}(|\xi_1|>t)}\,dt + 2pU\log\left(\frac{AU}{\delta}\right)\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] \\
&= 4L\left\|\xi\right\|_{2,1}\delta\sqrt{n}\sqrt{p\log\left(\frac{AU}{\delta}\right)} + 2pU\log\left(\frac{AU}{\delta}\right)\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] \qquad \text{(A.13)}
\end{aligned}$$

For the collection $\mathcal{F}_2$ we can directly use equation (A.11) for $k=n$ we obtain:

$$\mathbb{E}\left[\left\|\left(\mathbb{P}_n - P\right)f_{\theta,2}\right\|_{\mathcal{F}_2}\right] \leq L\left(\frac{\delta}{\sqrt{n}}\sqrt{p\log\frac{AU}{\delta}} + \frac{pU}{n}\log\frac{AU}{\delta}\right) \qquad \text{(A.14)}$$

Therefore combining equation (A.12) and (A.14) we conclude:

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0)\leq\delta}|(\mathbb{P}_n - P)\, f_\theta|\right] \leq L(4\,\|\xi\|_{2,1}+1)\frac{\delta}{\sqrt{n}}\sqrt{p\log\frac{AU}{\delta}} + \frac{pU}{n}\log\frac{AU}{\delta}\left(1+2\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right]\right) \tag{A.15}$$

Ignoring constants (as they won't effect the rate of convergence) we can take $\phi_n(\delta)$ is Theorem 3.4.1 of [38] as:

$$\phi_n(\delta) = \delta\sqrt{p\log\frac{1}{\delta}} \vee \frac{p}{\sqrt{n}}\log\frac{1}{\delta}\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right].$$

Finally solving the equation the equation $r_n^2\phi_n(1/r_n) \leq \sqrt{n}$ we conclude the rate of convergence.

Rate of convergence of the final estimators: The calculation is exactly same as in Paragraph A.4.1 and hence skipped.

A.5. Proof of Theorem 2.1

To establish the rate of convergence and the asymptotic distribution of the change point estimators obtained via Huber loss, we first need to establish a curvature of the population loss function around its unique minimizer. The following lemma is imperative to that end:

Lemma A.8. *If ξ follows a symmetric distribution around the origin with with continuous density f_ξ satisfying $f_\xi(0) > 0$, then for any $k > 0, |\mu| < 2k$, we have:*

$$\mathbb{E}\left[\tilde{H}_k(\xi+\mu) - \tilde{H}_k(\xi)\right] \geq \frac{\mu^2}{2}\mathbb{P}\left(-k \leq \xi \leq k-\mu\right) \geq \frac{\mu^2}{2}\mathbb{P}\left(-k \leq \xi \leq 0\right).$$

For $k = 0$, if we choose δ such that for all $|x| \leq \delta$, $f_\xi(x) \geq f_\xi(0)/2$, then we have for $|\mu| \leq \delta$:

$$\mathbb{E}\left[\tilde{H}_k(\xi+\mu) - \tilde{H}_k(\xi)\right] \geq \frac{\mu^2}{2}f_\xi(0).$$

The proof of the above lemma can be found in Appendix B. Now set $\delta > 0$ such that $\beta_0 + \delta < \alpha_0$. Define the empirical stochastic process $\mathbb{M}_n(\theta)$ as:

$$\begin{aligned}\mathbb{M}_n(\theta) &\equiv \mathbb{M}_n(\alpha,\beta,d)\\ &= \frac{1}{n}\sum_{i=1}^n \tilde{H}_k\left(Y_i - \alpha\mathbb{1}_{X_i\leq d} - \beta\mathbb{1}_{X_i>d}\right)\\ &= \frac{1}{n}\sum_{i=1}^n \tilde{H}_k\left(\xi_i + \alpha_0\mathbb{1}_{X_i\leq d_0} + \beta_0\mathbb{1}_{X_i>d_0} - \alpha\mathbb{1}_{X_i\leq d} - \beta\mathbb{1}_{X_i>d}\right)\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_k\left(\xi_i + \alpha_0 - \alpha\right) \mathbb{1}_{X_i \le d_0 \wedge d} + \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_k\left(\xi_i + \alpha_0 - \beta\right) \mathbb{1}_{d < X_i \le d_0}$$
$$+ \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_k\left(\xi_i + \beta_0 - \alpha\right) \mathbb{1}_{d_0 < X_i \le d} + \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_k\left(\xi_i + \beta_0 - \beta\right) \mathbb{1}_{X_i > d \vee d_0}$$

This implies the centred empirical stochastic process is:

$$\begin{aligned}
\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) &= \mathbb{E}\left[\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{H}_k\left(\xi_i + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{X_i \le d_0 \wedge d} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{H}_k\left(\xi_i + \alpha_0 - \beta\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{d < X_i \le d_0} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{H}_k\left(\xi_i + \beta_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{d_0 < X_i \le d} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{H}_k\left(\xi_i + \beta_0 - \beta\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{X_i > d \vee d_0}
\end{aligned}$$

and the corresponding population deterministic process:

$$\begin{aligned}
\mathbb{M}(\theta) - \mathbb{M}(\theta_0) &= \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X \le d \wedge d_0\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \alpha_0 - \beta\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(d < X < d_0\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \beta_0 - \alpha\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(d_0 < X < d\right) \\
&\quad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \beta_0 - \beta\right) - \tilde{H}_k(\xi_i)\right)\right] \mathbb{P}\left(X > d \vee d_0\right) \\
&\ge \frac{C_k}{2}\left[(\alpha_0 - \alpha)^2 \mathbb{P}\left(X \le d \wedge d_0\right) + (\beta_0 - \beta)^2 \mathbb{P}\left(X > d \vee d_0\right)\right. \\
&\qquad \left. + |d - d_0| \left\{2(\alpha_0 - \beta_0 - \delta)^2\right\}\right] \\
&\ge C_k\left[(\alpha_0 - \alpha)^2 + (\beta_0 - \beta)^2 + |d - d_0|\right] \qquad \text{(A.16)}
\end{aligned}$$

for all $|\alpha - \alpha_0| \le \delta, |\beta - \beta_0| \le \delta$, where the penultimate inequality follows from Lemma A.8. Also note that the definition of C_k is different in the last two lines, but as they are constant, we refrain ourselves from using different notations in each line.

Consistency: We have established that $\mathbb{M}(\theta)$ has local quadratic curvature with respect to (α, β) in a $\delta-$ neighbourhood around the truth. Now to establish the rate of convergence, we first need to establish the consistency of our estimator. To that end, we use Theorem 3.2.2 of [38]. That the process $\mathbb{M}(\theta) - \mathbb{M}(\theta_0)$ is continuous with respect to θ and has a clear minima at $\theta = \theta_0$ is immediate from the definition. Also the tightness of the minimizer $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{d})$ follows

directly from our assumption of the compact parameter space Θ. Therefore, all we need to show is that for any compact subset $K \subseteq \Theta$:

$$\sup_{\theta \in K} |(\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) - (\mathbb{M}(\theta) - \mathbb{M}(\theta_0))| = o_p(1)\,.$$

Towards that direction, define the function $f_\theta(X, \xi)$ as:

$$\begin{aligned} f_\theta(X,\xi) = &\left(\tilde{H}_k\left(\xi + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X \le d_0 \wedge d} \\ &\left(\tilde{H}_k\left(\xi + \alpha_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{d < X \le d_0} \\ &+ \left(\tilde{H}_k\left(\xi + \beta_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{d_0 < X \le d} \\ &+ \left(\tilde{H}_k\left(\xi + \beta_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X > d \vee d_0} \end{aligned}$$

It is immediate from the definition of $f_\theta(X, \xi)$ that the collection of functions:

$$\mathcal{F}_K = \{f_\theta : \theta \in K\}$$

has finite VC dimension. Furthermore, as K is compact, there exist c such that:

$$\max_{\theta \in K} \{|\alpha|, |\beta|, |d|\} \le c\,.$$

Note that the Huber function H_k is Lipschitz with the Lipschitz constant being k. Therefore we have for any $\mu > 0$:

$$\left|\tilde{H}_k(\xi + \mu) - \tilde{H}_k(\xi)\right| = \frac{k+1}{k} |H_k(\xi + \mu) - H_k(\xi)| \le k|\mu|\,. \tag{A.17}$$

This implies that the function of $\mathcal{F}$ are uniformly bounded. Hence using Glivenko-Cantelli theorem (e.g. see Theorem 2.8.1 of [38]) we conclude that:

$$\sup_{\theta \in K} |(\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) - (\mathbb{M}(\theta) - \mathbb{M}(\theta_0))| = \|(\mathbb{P}_n - P)\|_{\mathcal{F}} = o_p(1)\,.$$

This establishes the consistency.

Tightness upon proper scaling: We next show that:

$$\max\left\{\sqrt{n}\left(\hat{\alpha} - \alpha_0\right), \sqrt{n}\left(\hat{\beta} - \beta_0\right), n\left(\hat{d} - d_0\right)\right\} = O_p(1)\,.$$

Here we apply Theorem 3.2.5 of [38]. Define a semi-metric on Θ as:

$$d\left(\theta_1, \theta_2\right) = \sqrt{\left(\alpha_1 - \alpha_2\right)^2 + \left(\beta_1 - \beta_2\right)^2 + |d_1 - d_2|}$$

From (A.16) we have $\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \ge C_k d^2(\theta, \theta_0)$. To establish asymptotic equicontinuity of the process we need to bound:

$$\mathbb{E}\left[\sup_{d(\theta,\theta_0) \le \delta} |\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0) - (\mathbb{M}(\theta) - \mathbb{M}(\theta_0))|\right]$$

$$= \mathbb{E}\left[\sup_{d(\theta,\theta_0)\leq\delta} |(\mathbb{P}_n - P) f_\theta|\right]$$
$$= \mathbb{E}\left[\|\mathbb{P}_n - P\|_{\mathcal{F}_\delta}\right]$$

where we define the collection $\mathcal{F}_\delta$ as $\mathcal{F}_\delta = \{f_\theta : d(\theta,\theta_0) \leq \delta\}$. The envelope function of $\mathcal{F}_\delta$ is defined as:

$$\begin{aligned}
&\sup_{\theta:d(\theta,\theta_0)\leq\delta} \left|\left(\tilde{H}_k\left(\xi+\alpha_0-\alpha\right)-\tilde{H}_k(\xi)\right)\right| \mathbb{1}_{X\leq d_0\wedge d} \\
&\qquad \left|\left(\tilde{H}_k\left(\xi+\alpha_0-\beta\right)-\tilde{H}_k(\xi)\right)\right| \mathbb{1}_{d<X\leq d_0} \\
&\qquad + \left|\left(\tilde{H}_k\left(\xi+\beta_0-\alpha\right)-\tilde{H}_k(\xi)\right)\right| \mathbb{1}_{d_0<X\leq d} \\
&\qquad + \left|\left(\tilde{H}_k\left(\xi+\beta_0-\beta\right)-\tilde{H}_k(\xi)\right)\right| \mathbb{1}_{X>d\vee d_0} \\
&\leq C_k\left(2\delta + \mathbb{1}_{d<X\leq d_0} + \mathbb{1}_{d_0<X\leq d}\right) \\
&\leq 2C_k\left(2\delta \vee \mathbb{1}_{d<X\leq d_0} + \mathbb{1}_{d_0<X\leq d}\right) := F_\delta(X,\xi)
\end{aligned}$$

Hence the $L_2(P)$ norm of the envelope function:

$$\sqrt{PF_\delta^2} \leq 2C_k\left(2\delta \vee \sqrt{\mathbb{P}\left(d_0<X<d\right)+\mathbb{P}\left(d<X<d_0\right)}\right) \leq 4C_k\delta := \phi_n(\delta)\,.$$

Hence an application of Theorem 3.2.5 of [38] yields $\sqrt{n}\ d\left(\hat{\theta},\theta_0\right) = O_p(1)$, which completes the proof.

Asymptotic distribution: In the final paragraph we establish the asymptotic distribution of $\sqrt{n}\left(\hat{\alpha}-\alpha_0\right), \sqrt{n}\left(\hat{\beta}-\beta_0\right)$ and $n\left(\hat{d}-d_0\right)$. Towards that end, we largely follow the approach of Subsection 14.5.1 of [23]. For any $\mathbf{h} := (h_1,h_2,h_3) \in \mathbb{R}^3$ define a paramter vector $\theta_{n,\mathbf{h}} = \alpha_0 + \frac{h_1}{\sqrt{n}}, \beta_0 + \frac{h_2}{\sqrt{n}}, d_0 + \frac{h_3}{n}$. Define a stochastic process $\mathbb{Q}_n$ on $\mathbb{R}^3$ as:

$$\begin{aligned}
\mathbb{Q}_n(h_1,h_2,h_3) &= n \times \mathbb{P}_n\left(f_{\theta_{n,\mathbf{h}}} - f_{\theta_0}\right) \\
&= \sum_{i=1}^n \left(\tilde{H}_k\left(\xi_i + \frac{h_1}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{X_i\leq d_0\wedge d} \\
&+ \sum_{i=1}^n \left(\tilde{H}_k\left(\xi_i + \alpha_0 - \beta_0 - \frac{h_2}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{d_0+\frac{h_3}{n}<X_i\leq d_0} \\
&+ \sum_{i=1}^n \left(\tilde{H}_k\left(\xi_i + \beta_0 - \alpha_0 - \frac{h_1}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{d_0<X_i\leq d_0+\frac{h_3}{n}} \\
&\qquad + \sum_{i=1}^n \left(\tilde{H}_k\left(\xi_i + \frac{h_2}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right) \mathbb{1}_{X_i>d\vee d_0} \\
&:= \mathbb{Q}_{n,1}(\mathbf{h}) + \mathbb{Q}_{n,2}(\mathbf{h}) + \mathbb{Q}_{n,3}(\mathbf{h}) + \mathbb{Q}_{n,4}(\mathbf{h})\,.
\end{aligned}$$

It is immediate from the definition of $\mathbb{Q}_n(\mathbf{h})$ that:

$$\widehat{\mathbf{h}}_n := \left(\sqrt{n}\left(\hat{\alpha}-\alpha_0\right), \sqrt{n}\left(\hat{\beta}-\beta_0\right), n\left(\hat{d}-d_0\right)\right) = \text{mid argmin}_{\mathbf{h}\in\mathbb{R}^3}\mathbb{Q}_n(\mathbf{h})\,.$$

We next show that there exist a stochastic process $\mathbb{Q}$ on $\mathbb{R}^3$ such that for any compact rectangle $\mathbb{I} = I_1 \times I_2 \times I_3 \subset \mathbb{R}^3$:

$$\mathbb{Q}_n|_{\mathbb{I}} \overset{\mathscr{L}}{\Longrightarrow} \mathbb{Q}|_{\mathbb{I}}\,.$$

where the process $\mathbb{Q}$ is defined as:

$$\begin{aligned}\mathbb{Q}(\mathbf{h}) = &\left(h_1\sigma_k\sqrt{F_X(d_0)}\times Z_1 + \frac{h_1^2}{2}\mu_k F_X(d_0)\right)\\ &+\left(h_2\sigma_k\sqrt{\bar{F}_X(d_0)}\times Z_2 + \frac{h_2^2}{2}\mu_k \bar{F}_X(d_0)\right)\\ &+\text{CPP}\left(\tilde{H}_k\left(\xi_i+(\alpha_0-\beta_0)\right)-\tilde{H}_k(\xi_i), f_X(\theta_0)\right)\,.\end{aligned}$$

with $Z_1, Z_2 \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ and CPP is (as defined in the main paper) compound Poisson process. Note that the stochastic process $\mathbb{Q}_n$ is continuous with respect to its first two co-ordinates and cadlag (right continuous with left limit) with respect to its third co-ordinate. Hence to establish the convergence of $\{\mathbb{Q}_n|_{\mathbb{I}}\}_{n\in\mathbb{N}}$ we need to use Skorohod topology. We mainly use Theorem 13.5 of [7] to establish the convergence result. Towards that end, define:

$$\begin{aligned}\tilde{\xi}_{i,h_1} &= \left(\tilde{H}_k\left(\xi_i+\frac{h_1}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right) - \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i+\frac{h_1}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right)\right]\\ \tilde{\xi}_{i,h_2} &= \left(\tilde{H}_k\left(\xi_i+\frac{h_2}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right) - \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i+\frac{h_2}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right)\right]\end{aligned}$$

and another stochastic process $\tilde{\mathbb{Q}}_n(\mathbf{h})$ as:

$$\begin{aligned}\tilde{\mathbb{Q}}_n(\mathbf{h}) = &\sum_{i=1}^n \tilde{\xi}_{i,h_1}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}}\\ &+\sum_{i=1}^n\left(\tilde{H}_k\left(\xi_i+(\alpha_0-\beta_0)\right)-\tilde{H}_k(\xi_i)\right)\mathbb{1}_{d_0+\frac{h_3}{n}<X_i\le d_0}\\ &+\sum_{i=1}^n\left(\tilde{H}_k\left(\xi_i+(\beta_0-\alpha_0)\right)-\tilde{H}_k(\xi_i)\right)\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}\\ &+\sum_{i=1}^n \tilde{\xi}_{i,h_2}\mathbb{1}_{X_i>d_0\vee d_0+\frac{h_3}{n}}\\ :=\ &\tilde{\mathbb{Q}}_{n,+}(\mathbf{h})\mathbb{1}_{h_3\ge 0}+\tilde{\mathbb{Q}}_{n,-}(\mathbf{h})\mathbb{1}_{h_3<0}\,.\end{aligned}$$

Hence we have the following decomposition:

$$\mathbb{Q}_n(\mathbf{h}) = \tilde{\mathbb{Q}}_n(\mathbf{h}) + \mathfrak{E}_n(\mathbf{h}) + \mathfrak{R}_n(\mathbf{h}) \tag{A.18}$$

where:

$$\mathfrak{R}_n(\mathbf{h}) = \sum_{i=1}^{n} \left(\tilde{H}_k \left(\epsilon_i + (\alpha_0 - \beta_0) - \frac{h_2}{\sqrt{n}} \right) - \tilde{H}_k \left(\epsilon_i + (\alpha_0 - \beta_0) \right) \right) \mathbb{1}_{d_0 + \frac{h_3}{n} < X_i \leq d_0} + \sum_{i=1}^{n} \left(\tilde{H}_k \left(\epsilon_i + (\beta_0 - \alpha_0) - \frac{h_1}{\sqrt{n}} \right) - \tilde{H}_k \left(\epsilon_i + (\beta_0 - \alpha_0) \right) \right) \mathbb{1}_{d_0 < X_i \leq d_0 + \frac{h_3}{n}}$$

and

$$\mathfrak{E}_n(\mathbf{h}) = \mathbb{E} \left[\left(\tilde{H}_k \left(\xi_i + \frac{h_1}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i \right) \right) \right] \sum_{i=1}^{n} \mathbb{1}_{X_i \leq d_0 \wedge d_0 + \frac{h_3}{n}} + \mathbb{E} \left[\left(\tilde{H}_k \left(\xi_i + \frac{h_2}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i \right) \right) \right] \sum_{i=1}^{n} \mathbb{1}_{X_i > d_0 \vee d_0 + \frac{h_3}{n}}$$

We next show $\mathfrak{R}_n(\mathbf{h})$ is $o_p(1)$ uniformly over a compact set:

$$\begin{aligned}
&|\mathfrak{R}_n(\mathbf{h})| \\
&= \left| \sum_{i=1}^{n} \left(\tilde{H}_k \left(\xi_i + (\alpha_0 - \beta_0) - \frac{h_2}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i + (\alpha_0 - \beta_0) \right) \right) \mathbb{1}_{d_0 + \frac{h_3}{n} < X_i \leq d_0} \right. \\
&\left. + \sum_{i=1}^{n} \left(\tilde{H}_k \left(\xi_i + (\beta_0 - \alpha_0) - \frac{h_1}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i + (\beta_0 - \alpha_0) \right) \right) \mathbb{1}_{d_0 < X_i \leq d_0 + \frac{h_3}{n}} \right| \\
&\leq \sum_{i=1}^{n} \left| \tilde{H}_k \left(\xi_i + (\alpha_0 - \beta_0) - \frac{h_2}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i + (\alpha_0 - \beta_0) \right) \right| \mathbb{1}_{d_0 + \frac{h_3}{n} < X_i \leq d_0} \\
&+ \sum_{i=1}^{n} \left| \tilde{H}_k \left(\xi_i + (\beta_0 - \alpha_0) - \frac{h_1}{\sqrt{n}} \right) - \tilde{H}_k \left(\xi_i + (\beta_0 - \alpha_0) \right) \right| \mathbb{1}_{d_0 < X_i \leq d_0 + \frac{h_3}{n}} \\
&\leq \frac{(k+1)h_2}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}_{d_0 + \frac{h_3}{n} < X_i \leq d_0} + \frac{(k+1)h_1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}_{d_0 < X_i \leq d_0 + \frac{h_3}{n}}
\end{aligned}$$

Now suppose $\mathbf{h} \in \mathbb{I}$. There $|h_1| \vee |h_2| \vee |h_3| \leq K$ for some $K > 0$. Hence we have:

$$\begin{aligned}
\sup_{\mathbf{h} \in \mathbb{I}} |\mathfrak{R}_n(\mathbf{h})| &\leq \frac{(k+1)K}{\sqrt{n}} \left[\sum_{i=1}^{n} \mathbb{1}_{d_0 - \frac{K}{n} < X_i \leq d_0} + \sum_{i=1}^{n} \mathbb{1}_{d_0 < X_i \leq d_0 + \frac{K}{n}} \right] \\
&= \frac{(k+1)K}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{1}_{d_0 - \frac{K}{n} < X_i \leq d_0 + \frac{K}{n}}
\end{aligned}$$

as we know:

$$\sum_{i=1}^{n} \mathbb{1}_{d_0 - \frac{K}{n} < X_i \leq d_0 + \frac{K}{n}} \overset{\mathscr{L}}{\Longrightarrow} \text{Pois} \left(K f_X(d_0) \right)$$

we conclude that:

$$\sup_{\mathbf{h} \in \mathbb{I}} |\mathfrak{R}_n(\mathbf{h})| = O_p(n^{-1/2}) = o_p(1) \,. \tag{A.19}$$

We now establish the convergence of $\mathfrak{E}(\mathbf{h})$:

$$
\begin{aligned}
\mathfrak{E}_n(\mathbf{h}) &= \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i+\frac{h_1}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right)\right]\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}} \\
&\qquad + \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i+\frac{h_2}{\sqrt{n}}\right)-\tilde{H}_k(\xi_i)\right)\right]\sum_{i=1}^{n}\mathbb{1}_{X_i> d_0\vee d_0+\frac{h_3}{n}} \\
&= \frac{k+1}{k}\left\{\frac{h_1^2}{2n}\mathbb{P}\left(-k\le\xi\le k-\frac{h_1}{\sqrt{n}}\right)\right. \\
&\quad +\mathbb{E}\left[\left(\frac{h_1}{\sqrt{n}}k-\frac{h_1}{\sqrt{n}}\xi-\frac{1}{2}(\xi-k)^2\right)\mathbb{1}_{k-\frac{h_1}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad \left.+\ \mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_1}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_1}{\sqrt{n}}\le\xi\le -k}\right]\right\}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}} \\
&\quad + \frac{k+1}{k}\left\{\frac{h_2^2}{2n}\mathbb{P}\left(-k\le\xi\le k-\frac{h_2}{\sqrt{n}}\right)\right. \\
&\quad +\mathbb{E}\left[\left(\frac{h_2}{\sqrt{n}}k-\frac{h_2}{\sqrt{n}}\xi-\frac{1}{2}(\xi-k)^2\right)\mathbb{1}_{k-\frac{h_2}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad \left.+\ \mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_2}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_2}{\sqrt{n}}\le\xi\le -k}\right]\right\}\sum_{i=1}^{n}\mathbb{1}_{X_i> d_0\vee d_0+\frac{h_3}{n}} \\
&= n\frac{k+1}{k}\left\{\frac{h_1^2}{2n}\mathbb{P}\left(-k\le\xi\le k-\frac{h_1}{\sqrt{n}}\right)\right. \\
&\quad +\mathbb{E}\left[\left(\frac{h_1}{\sqrt{n}}k-\frac{h_1}{\sqrt{n}}\xi-\frac{1}{2}(\xi-k)^2\right)\mathbb{1}_{k-\frac{h_1}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad \left.+\ \mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_1}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_1}{\sqrt{n}}\le\xi\le -k}\right]\right\}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}} \\
&\quad + n\frac{k+1}{k}\left\{\frac{h_2^2}{2n}\mathbb{P}\left(-k\le\xi\le k-\frac{h_2}{\sqrt{n}}\right)\right. \\
&\quad +\mathbb{E}\left[\left(\frac{h_2}{\sqrt{n}}k-\frac{h_2}{\sqrt{n}}\xi-\frac{1}{2}(\xi-k)^2\right)\mathbb{1}_{k-\frac{h_2}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad \left.+\ \mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_2}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_2}{\sqrt{n}}\le\xi\le -k}\right]\right\}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i> d_0\vee d_0+\frac{h_3}{n}} \\
&= \frac{k+1}{k}\left\{\frac{h_1^2}{2}\mathbb{P}\left(-k\le\xi\le k-\frac{h_1}{\sqrt{n}}\right)\right. \\
&\quad +n\mathbb{E}\left[\left(\frac{h_1}{\sqrt{n}}k-\frac{h_1}{\sqrt{n}}\xi-\frac{1}{2}(\xi-k)^2\right)\mathbb{1}_{k-\frac{h_1}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad \left.+\ n\mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_1}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_1}{\sqrt{n}}\le\xi\le -k}\right]\right\}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{k+1}{k}\left\{\frac{h_2^2}{2}\mathbb{P}\left(-k\le\xi\le k-\frac{h_2}{\sqrt{n}}\right)\right.\\
&+n\mathbb{E}\left[\left(\frac{h_2}{\sqrt{n}}k-\frac{h_2}{\sqrt{n}}\xi-\frac{1}{2}\left(\xi-k\right)^2\right)\mathbb{1}_{k-\frac{h_2}{\sqrt{n}}\le\xi\le k}\right]\\
&\left.+\ n\mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_2}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_2}{\sqrt{n}}\le\xi\le-k}\right]\right\}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0\vee d_0+\frac{h_3}{n}}
\end{aligned}
$$

From strong law of large numbers we have:

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0\wedge d_0+\frac{h_3}{n}}\xrightarrow{p}\mathbb{P}(X\le d_0)\,, \tag{A.20}
$$

$$
\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0\vee d_0+\frac{h_3}{n}}\xrightarrow{p}\mathbb{P}(X>d_0)\,, \tag{A.21}
$$

For the other terms in the expectation:

$$
\frac{h_1^2}{2}\mathbb{P}\left(-k\le\xi\le k-\frac{h_1}{\sqrt{n}}\right)\longrightarrow\frac{h_1^2}{2}\mathbb{P}\left(-k\le\epsilon\le k\right)\,, \tag{A.22}
$$

$$
\frac{h_2^2}{2}\mathbb{P}\left(-k\le\xi\le k-\frac{h_2}{\sqrt{n}}\right)\longrightarrow\frac{h_2^2}{2}\mathbb{P}\left(-k\le\epsilon\le k\right)\,. \tag{A.23}
$$

For the other terms:

$$
\begin{aligned}
&n\mathbb{E}\left[\left(\frac{h_2}{\sqrt{n}}k-\frac{h_2}{\sqrt{n}}\xi-\frac{1}{2}\left(\xi-k\right)^2\right)\mathbb{1}_{k-\frac{h_2}{\sqrt{n}}\le\xi\le k}\right]\\
&=n\int_{k-\frac{h_2}{\sqrt{n}}}^{k}\left(\frac{h_2}{\sqrt{n}}k-\frac{h_2}{\sqrt{n}}x-\frac{1}{2}\left(x-k\right)^2\right)f_\xi(x)\,dx\\
&=n\left[\frac{h_2}{\sqrt{n}}\int_{k-\frac{h_2}{\sqrt{n}}}^{k}(k-x)f_\xi(x)\,dx-\frac{1}{2}\int_{k-\frac{h_2}{\sqrt{n}}}^{k}\left(x-k\right)^2\,f_\xi(x)\,dx\right]\\
&=n\left[\frac{h_2}{\sqrt{n}}\int_{0}^{\frac{h_2}{\sqrt{n}}}zf_\xi(z-k)\,dx-\frac{1}{2}\int_{-\frac{h_2}{\sqrt{n}}}^{0}z^2\,f_\xi(z+k)\,dx\right]\\
&\le Cn\left[\frac{h_2}{\sqrt{n}}\int_{0}^{\frac{h_2}{\sqrt{n}}}z\,dz+\frac{1}{2}\int_{-\frac{h_2}{\sqrt{n}}}^{0}z^2\,dz\right]\qquad[C=\max_x f_\xi(x)]\\
&=n\times O(n^{-3/2})=o(1)\,.
\end{aligned} \tag{A.24}
$$

and

$$
\begin{aligned}
&n\mathbb{E}\left[\frac{1}{2}\left(\xi+\frac{h_2}{\sqrt{n}}+k\right)^2\mathbb{1}_{-k-\frac{h_2}{\sqrt{n}}\le\xi\le-k}\right]\\
&=n\int_{-k-\frac{h_2}{\sqrt{n}}}^{-k}\frac{1}{2}\left(x+\frac{h_2}{\sqrt{n}}+k\right)^2\,f_\xi(x)\,dx
\end{aligned}
$$

$$= \frac{h_2^3}{6n^{1/2}} f_\xi(k) + o(1) = o(1)\,. \tag{A.25}$$

Similar calculation holds for the terms involving h_1. Hence we conclude combining equations (A.20) - (A.25) we conclude:

$$\mathfrak{E}_n(\mathbf{h}) \xrightarrow{P} \frac{k+1}{k}\left[\frac{h_1^2}{2}\mathbb{P}\left(-k \le \epsilon \le k\right) F_X(d_0) + \frac{h_2^2}{2}\mathbb{P}\left(-k \le \epsilon \le k\right) \bar{F}_X(d_0)\right]. \tag{A.26}$$

Finally we show the weak convergence of $\tilde{\mathbb{Q}}_n(\mathbf{h})$ to $\mathbb{Q}(\mathbf{h})$, for which we need to show that the collection $\{\tilde{\mathbb{Q}}_n(\mathbf{h})\}_{n\in\mathbb{N}}$ is tight with respect to appropriate topology and every finite dimensional projection of $\tilde{\mathbb{Q}}_n(\mathbf{h})$ converges to that of $\mathbb{Q}(\mathbf{h})$. We embark on by showing that for any fixed $\mathbf{h}$, $\tilde{\mathbb{Q}}_n(\mathbf{h})$ converges to $\mathbb{Q}(\mathbf{h})$ in distribution. Towards that direction, fix $h_3 > 0$:

$$\begin{aligned}\mathbb{E}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right) \mid X_1, \ldots, X_n\right]\right] \\ &:= \mathbb{E}\left[\mathbb{E}_{\mathbf{X}}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right]\right]\end{aligned}$$

We start with analyzing the inner expectation $\mathbb{E}_{\mathbf{X}}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right]$:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right] &= \mathbb{E}_{\mathbf{X}}\left[\exp\left\{\left(it\sum_{i=1}^{n}\tilde{\xi}_{i,h_1}\mathbb{1}_{X_i\le d_0}\right.\right.\right. \\ &\qquad + \sum_{i=1}^{n}\left(\tilde{H}_k\left(\xi_i + (\alpha_0-\beta_0)\right) - \tilde{H}_k(\xi_i)\right)\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}} \\ &\qquad \left.\left.\left. + \sum_{i=1}^{n}\tilde{\xi}_{i,h_2}\mathbb{1}_{X_i>d_0+\frac{h_3}{n}}\right)\right\}\right] \\ &= \left(\phi_{\tilde{\xi}_{h_1}}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0}} \times \left(\phi_{\tilde{\xi}_{h_2}}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0+\frac{h_3}{n}}} \\ &\qquad \times \left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}} \\ &= \left(\phi_{\tilde{\xi}_{h_1}}(t)\right)^{n\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0}} \times \left(\phi_{\tilde{\xi}_{h_2}}(t)\right)^{n\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0+\frac{h_3}{n}}} \\ &\qquad \times \left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}}\end{aligned} \tag{A.27}$$

To show the convergence of the characteristic function of $\tilde{\xi}_{h_1}$ (and similarly for $\tilde{\xi}_{h_2}$) we first note that the variance of $\tilde{\xi}_{h_1}$ for $h_1 > 0$ is:

$$\begin{aligned}\mathsf{var}\left(\tilde{\xi}\right) &= \mathsf{var}\left(\tilde{H}_k\left(\xi_i + \frac{h_1}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right) \\ &= \mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i + \frac{h_1}{\sqrt{n}}\right) - \tilde{H}_k(\xi_i)\right)^2\right] + O(n^{-2})\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{k+1}{k}\right)^2 \Bigg\{ \mathbb{E}\left[\left(\frac{h_1}{\sqrt{n}}\xi + \frac{h_1^2}{2n}\right)^2 \mathbb{1}_{-k\le\xi\le k-\frac{h_1}{\sqrt{n}}}\right] \\
&\quad + \mathbb{E}\left[\left(\frac{1}{2}\left(\xi + \frac{h_1}{\sqrt{n}} + k\right)^2 - \frac{h_1}{\sqrt{n}}k\right)^2 \mathbb{1}_{-k-\frac{h_1}{\sqrt{n}}\le\xi\le -k}\right] \\
&\quad + \mathbb{E}\left[\left(-\frac{1}{2}(\xi-k)^2 + k\frac{h_1}{\sqrt{n}}\right)^2 \mathbb{1}_{k-\frac{h_1}{\sqrt{n}}\le\xi\le k}\right] \\
&\quad + \frac{k^2h_1^2}{n}\left\{\mathbb{P}(\xi > k) + \mathbb{P}\left(\xi < -k - \frac{h_1}{\sqrt{n}}\right)\right\} + O(n^{-2})\Bigg\} \\
&= \left(\frac{k+1}{k}\right)^2 \left(\frac{h_1^2}{n}\mathbb{E}\left[\xi^2\mathbb{1}_{-k\le\xi\le k}\right] + \frac{k^2h_1^2}{n}\left\{\mathbb{P}(\xi > k) + \mathbb{P}(\xi < -k)\right\}\right) + o(n^{-1}) \\
&= \left(\frac{k+1}{k}\right)^2 \left(\frac{h_1^2}{n}\mathbb{E}\left[\xi^2\mathbb{1}_{-k\le\xi\le k}\right] + \frac{2k^2h_1^2}{n}\mathbb{P}(\xi > k)\right) + o(n^{-1}) \\
&:= \frac{h_1^2\sigma_k^2}{n} + o(n^{-1}).
\end{aligned}
$$

where the variance parameter σ_k^2 is defined as:

$$
\left(\frac{k+1}{k}\right)^2 \left(\mathbb{E}\left[\xi^2\mathbb{1}_{-k\le\xi\le k}\right] + 2k^2\mathbb{P}(\xi > k)\right).
$$

Similar calculation holds for $h_1 < 0$ and for h_2. Hence going back to equation (A.27) we have:

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right] &= \left(1 - \frac{t^2}{2}\frac{h_1^2\sigma_k^2}{n} + o(n^{-1})\right)^{n\times\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\le d_0}} \\
&\quad \times \left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\sum_{i=1}^n \mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}} \\
&\quad \times \left(1 - \frac{t^2}{2}\frac{h_2^2\sigma_k^2}{n} + o(n^{-1})\right)^{n\times\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i>d_0+\frac{h_3}{n}}}
\end{aligned}
$$

As

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\le d_0} &\xrightarrow{a.s.} F_X(d_0) \\
\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i>d_0+\frac{h_3}{n}} &\xrightarrow{a.s.} \bar{F}_X(d_0)
\end{aligned}
$$

we conclude:

$$
\left(1 - \frac{t^2}{2}\frac{h_1^2\sigma_k^2}{n} + o(n^{-1})\right)^{n\times\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\le d_0}} \xrightarrow{a.s.} e^{-\frac{t^2h_1^2\sigma_k^2}{2}F_X(d_0)}
$$

$$\left(1-\frac{t^2}{2}\frac{h_2^2\sigma_k^2}{n}+o(n^{-1})\right)^{n\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0+\frac{h_3}{n}}} \xrightarrow{a.s.} e^{-\frac{t^2h_2^2\sigma_k^2}{2}\bar{F}_X(d_0)}$$

Also we know:

$$\sum_{i=1}^{n}\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}} \overset{\mathscr{L}}{\Longrightarrow} \text{Pois}\left(f_X(d_0)h_3\right)$$

which further implies:

$$\begin{aligned}&\left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}}\\ &\overset{\mathscr{L}}{\Longrightarrow}\left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\text{Pois}(f_X(\theta_0)h_3)}\end{aligned}$$

Hence combining these we conclude:

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right] &= \left(\phi_{\tilde{\xi}_{h_1}}(t)\right)^{n\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i\le d_0}}\times\left(\phi_{\tilde{\xi}_{h_2}}(t)\right)^{n\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i>d_0+\frac{h_3}{n}}}\\ &\quad\times\left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\sum_{i=1}^{n}\mathbb{1}_{d_0<X_i\le d_0+\frac{h_3}{n}}}\\ &\overset{\mathscr{L}}{\Longrightarrow} e^{-\frac{t^2h_1^2\sigma_k^2}{2}F_X(d_0)}\times e^{-\frac{t^2h_2^2\sigma_k^2}{2}\bar{F}_X(d_0)}\\ &\quad\times\left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\text{Pois}(f_X(d_0)h_3)}\end{aligned}$$

Applying DCT and taking expectation on the both side we conclude:

$$\begin{aligned}\mathbb{E}\left[\exp\left(it\tilde{\mathbb{Q}}_n(\mathbf{h})\right)\right] &\longrightarrow e^{-\frac{t^2h_1^2\sigma_k^2}{2}F_X(d_0)}\times e^{-\frac{t^2h_2^2\sigma_k^2}{2}\bar{F}_X(d_0)}\\ &\quad\times\mathbb{E}\left[\left(\phi_{\tilde{H}_k(\xi_i+(\alpha_0-\beta_0))-\tilde{H}_k(\xi_i)}(t)\right)^{\text{Pois}(f_X(d_0)h_3)}\right].\end{aligned}$$

This concludes that $\tilde{\mathbb{Q}}_n(\mathbf{h}) \overset{\mathscr{L}}{\Longrightarrow} \mathbb{Q}(\mathbf{h})$. The proof of the fact that for any finite collection $(\mathbf{h}_1,\ldots,\mathbf{h}_l)$:

$$\left(\tilde{\mathbb{Q}}_n(\mathbf{h}_1),\ldots,\tilde{\mathbb{Q}}_n(\mathbf{h}_l)\right) \overset{\mathscr{L}}{\Longrightarrow} \left(\mathbb{Q}(\mathbf{h}_1),\ldots,\mathbb{Q}(\mathbf{h}_l)\right) \tag{A.28}$$

is similar (same analysis of characteristic function) and hence skipped for brevity. Interested readers can take a look at the proof of Lemma 3.2 of [26] or the proof of Theorem 5 of [24] for more details of this type of calculations. We next establish the tightness of the process. Define another process $\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h})$ as:

$$\begin{aligned}\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h}) &= \sum_{i=1}^{n}\tilde{\xi}_{i,h_1}\mathbb{1}_{X_i\le d_0}\\ &\quad+\sum_{i=1}^{n}\left(H_k\left(\xi_i+(\alpha_0-\beta_0)\right)-\tilde{H}_k(\xi_i)\right)\mathbb{1}_{d_0+\frac{h_3}{n}<X_i\le d_0}\end{aligned}$$

$$+\sum_{i=1}^{n}\left(H_k\left(\xi_i+(\alpha_0-\beta_0)\right)-\tilde{H}_k(\xi_i)\right)\mathbb{1}_{d_0<X_i\leq d_0+\frac{h_3}{n}}$$
$$+\sum_{i=1}^{n}\tilde{\xi}_{i,h_2}\mathbb{1}_{X_i>d_0}$$
$$:=\tilde{\tilde{\mathbb{Q}}}_{n,1}(\mathbf{h})+\tilde{\tilde{\mathbb{Q}}}_{n,2}(\mathbf{h})+\tilde{\tilde{\mathbb{Q}}}_{n,3}(\mathbf{h})+\tilde{\tilde{\mathbb{Q}}}_{n,4}(\mathbf{h}) \tag{A.29}$$

We now show that $\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h})$ uniformly approximate $\tilde{\mathbb{Q}}_n(\mathbf{h})$ over compact sets:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\mathbf{h}\in\mathbb{I}}\left|\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h})-\tilde{\mathbb{Q}}_n(\mathbf{h})\right|\right]\\
&\leq\mathbb{E}\left[\sup_{\mathbf{h}\in\mathbb{I}}\left\{\left|\sum_{i=1}^{n}\tilde{\xi}_{i,h_1}\left[\mathbb{1}_{X_i\leq d_0}-\mathbb{1}_{X_i\leq d_0\wedge d_0+\frac{h_3}{n}}\right]\right|\right.\right.\\
&\qquad\left.\left.+\left|\sum_{i=1}^{n}\tilde{\xi}_{i,h_2}\left[\mathbb{1}_{X_i>d_0}-\mathbb{1}_{X_i>d_0\vee d_0+\frac{h_3}{n}}\right]\right|\right\}\right]\\
&\leq 2\mathbb{E}\left[\sup_{\mathbf{h}\in\mathbb{I}}\left|\sum_{i=1}^{n}\tilde{\xi}_{i,h_1}\left[\mathbb{1}_{X_i\leq d_0}-\mathbb{1}_{X_i\leq d_0\wedge d_0+\frac{h_3}{n}}\right]\right|\right]\\
&\qquad+2\mathbb{E}\left[\sup_{\mathbf{h}\in\mathbb{I}}\left|\sum_{i=1}^{n}\tilde{\xi}_{i,h_2}\left[\mathbb{1}_{X_i\leq d_0}-\mathbb{1}_{X_i>d_0\vee d_0+\frac{h_3}{n}}\right]\right|\right]\\
&\leq 2\mathbb{E}\left[\sum_{i=1}^{n}|\tilde{\xi}_{i,h_1}|\mathbb{1}_{d_0-\frac{K}{n}<X_i\leq d_0}\right]+2\mathbb{E}\left[\sum_{i=1}^{n}|\tilde{\xi}_{i,h_2}|\mathbb{1}_{d_0<X_i\leq d_0+\frac{K}{n}}\right]\\
&\leq 2n\mathbb{E}\left[|\tilde{\xi}_{h_1}|\right]\mathbb{P}\left(d_0-\frac{K}{n}<X\leq d_0\right)+2n\mathbb{E}\left[|\tilde{\xi}_{h_2}|\right]\mathbb{P}\left(d_0<X\leq d_0+\frac{K}{n}\right)\\
&\leq 2n\times\left[\sqrt{\mathsf{var}\left(\tilde{\xi}_{h_1}\right)}\times\left(\frac{K}{n}f_X(d_0)+o(n^{-1})\right)\right.\qquad\text{(A.30)}\\
&\qquad\left.+\sqrt{\mathsf{var}\left(\tilde{\xi}_{h_2}\right)}\times\left(\frac{K}{n}f_X(d_0)+o(n^{-1})\right)\right]\\
&=2n\times\left[\left(\frac{h_1\sigma_k}{\sqrt{n}}+o(n^{-1/2})\right)\times\left(\frac{K}{n}f_X(d_0)+o(n^{-1})\right)\right.\\
&\qquad\left.+\left(\frac{h_2\sigma_k}{\sqrt{n}}+o(n^{-1/2})\right)\times\left(\frac{K}{n}f_X(d_0)+o(n^{-1})\right)\right]\\
&=O(n^{-1/2})=o(1)\,.\qquad\text{(A.31)}
\end{aligned}$$

Hence from equation (A.31) we conclude:

$$\sup_{\mathbf{h}\in\mathbb{I}}\left|\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h})-\tilde{\mathbb{Q}}_n(\mathbf{h})\right|=o_p(1)\,. \tag{A.32}$$

Therefore it is immediate from equation (A.28) that:

$$\left(\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h}_1),\ldots,\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h}_l)\right)\overset{\mathscr{L}}{\Longrightarrow}\left(\mathbb{Q}(\mathbf{h}_1),\ldots,\mathbb{Q}(\mathbf{h}_l)\right)\,. \tag{A.33}$$

Next, we show that tightness $\left\{\tilde{\tilde{\mathbb{Q}}}_n(\mathbf{h})\right\}_{n\in\mathbb{N}}$. As evident from equation (A.29), it is enough to show tightness of $\left\{\tilde{\tilde{\mathbb{Q}}}_{n,i}(\mathbf{h})\right\}_{n\in\mathbb{N}}$ for $i = 1, 2, 3, 4$. For $i = 1$, the process $\tilde{\tilde{\mathbb{Q}}}_{n,1}(\mathbf{h})$ only depends on h_1 and hence have continuous paths. Therefore to establish tightness, it is enough to show:

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty}\mathbb{E}\left[\sup_{\substack{|h_{1,1}|\vee|h_{1,2}|\le K\\ |h_{1,1}-h_{1,2}|\le\delta}}\sum_{i=1}^{n}\left|\tilde{\xi}_{i,h_{1,1}}-\tilde{\xi}_{i,h_{1,2}}\right|\mathbb{1}_{X_i\le d_0}\right]=0$$

Towards that end, define a collection of functions:

$$\mathcal{F}_{1,\delta}=\left\{f_{h_{1,1},h_{1,2}}:|h_{1,1}|\vee|h_{1,2}|\le K,|h_{1,1}-h_{1,2}|\le\delta\right\},$$

where the individual functions $f_{h_{1,1},h_{1,2}}$ is defined as:

$$f_{h_{1,1},h_{1,2}}(X,\epsilon)=\left\{\left(\tilde{H}_k\left(\xi_i+\frac{h_1}{\sqrt{n}}\right)-\tilde{H}_k\left(\xi_i+\frac{h_2}{\sqrt{n}}\right)\right)\right.$$
$$\left.-\mathbb{E}\left[\left(\tilde{H}_k\left(\xi_i+\frac{h_1}{\sqrt{n}}\right)-\tilde{H}_k\left(\xi_i+\frac{h_2}{\sqrt{n}}\right)\right)\right]\right\}\mathbb{1}_{X\le d_0}.$$

Clearly $\mathcal{F}_{1,\delta}$ has finite VC dimension. Also from equation (A.17) we have:

$$\begin{aligned}&\left|\tilde{H}_k\left(\xi_i+\frac{h_{1,1}}{\sqrt{n}}\right)-\tilde{H}_k\left(\xi_i+\frac{h_{1,2}}{\sqrt{n}}\right)\right|\\&\le\frac{k+1}{k}\left[4k\frac{|h_{1,1}-h_{1,2}|}{\sqrt{n}}+\frac{1}{2}\frac{(h_{1,1}-h_{1,2})^2}{n}+\left(\frac{|h_{1,1}-h_{1,2}|}{\sqrt{n}}\wedge 2k\right)^2\right]\\&\le C_k\frac{\delta}{\sqrt{n}}.\end{aligned}$$

Therefore, the envelope function can be taken as:

$$F_{1,\delta}(X,\epsilon)=\frac{2C_k\delta}{\sqrt{n}}.$$

Hence using Lemma 2.14.1 of [38] we conclude:

$$\limsup_{n\to\infty}\mathbb{E}\left[\sup_{\substack{|h_{1,1}|\vee|h_{1,2}|\le K\\ |h_{1,1}-h_{1,2}|\le\delta}}\sum_{i=1}^{n}\left|\tilde{\xi}_{i,h_{1,1}}-\tilde{\xi}_{i,h_{1,2}}\right|\mathbb{1}_{X_i\le d_0}\right]\lesssim\delta$$

which established tightness of $\tilde{\tilde{\mathbb{Q}}}_{n,1}(\mathbf{h})$. The proof of tightness of $\tilde{\tilde{\mathbb{Q}}}_{n,4}(\mathbf{h})$ is similar and hence skipped. Finally we show the tightness of $\tilde{\tilde{\mathbb{Q}}}_{n,23}(\mathbf{h})=\tilde{\tilde{\mathbb{Q}}}_{n,2}(\mathbf{h})+\tilde{\tilde{\mathbb{Q}}}_{n,3}(\mathbf{h})$. As these terms only depend on h_3 which has cadlag paths, we use equation (13.14) of Theorem 13.5 from [7] with $\beta = 1/2$, $\alpha = 1$ and $F(x) = Cx$

for some constant C. Fix $h_{3,1} < 0 < h_{3,2} < h_{3,3}$. The other cases (i.e. say $0 < h_{3,1} < h_{3,2} < h_{3,3}$) are similar and hence skipped.

$$
\begin{aligned}
&\mathbb{E}\left[\left|\tilde{\tilde{\mathbb{Q}}}_{n,23}(h_{3,1}) - \tilde{\tilde{\mathbb{Q}}}_{n,23}(h_{3,2})\right| \left|\tilde{\tilde{\mathbb{Q}}}_{n,23}(h_{3,2}) - \tilde{\tilde{\mathbb{Q}}}_{n,23}(h_{3,3})\right|\right] \\
&= \mathbb{E}\left[\left|\sum_{i=1}^{n}\left(\tilde{H}_k\left(\xi_i + (\alpha_0 - \beta_0)\right) - \tilde{H}_k(\xi_i)\right)\left[\mathbb{1}_{d_0 + \frac{h_{3,1}}{n} \le X_i < d_0} - \mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,2}}{n}}\right]\right|\right. \\
&\times \left.\left|\sum_{i=1}^{n}\left(\tilde{H}_k\left(\xi_i + (\alpha_0 - \beta_0)\right) - \tilde{H}_k(\xi_i)\right)\left[\mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,2}}{n}} - \mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,3}}{n}}\right]\right|\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\left|\tilde{H}_k\left(\xi_i + (\alpha_0 - \beta_0)\right) - \tilde{H}_k(\xi_i)\right|\left[\mathbb{1}_{d_0 + \frac{h_{3,1}}{n} \le X_i < d_0} + \mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,2}}{n}}\right]\right. \\
&\times \left.\sum_{i=1}^{n}\left|\tilde{H}_k\left(\xi_i + (\alpha_0 - \beta_0)\right) - \tilde{H}_k(\xi_i)\right|\left[\mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,2}}{n}} + \mathbb{1}_{d_0 \le X_i < d_0 + \frac{h_{3,3}}{n}}\right]\right] \\
&\le \frac{k+1}{k}\left(4(\alpha_0 - \beta_0)k + \frac{1}{2}(\alpha_0 - \beta_0)^2 + \left((\alpha_0 - \beta_0) \wedge 2k\right)^2\right) \times \\
&\qquad\qquad \mathbb{E}\left[\left(\sum_{i=1}^{n}\mathbb{1}_{d_0 + \frac{h_{3,1}}{\sqrt{n}} \le X_i < d_0 + \frac{h_{3,2}}{n}}\right) \times \left(\sum_{i=1}^{n}\mathbb{1}_{d_0 + \frac{h_{3,2}}{n} \le X_i < d_0 + \frac{h_{3,3}}{n}}\right)\right] \\
&= C_{k,\theta_0}\sum_{i \ne j}\mathbb{E}\left[\mathbb{1}_{d_0 + \frac{h_{3,1}}{n} \le X_i < d_0 + \frac{h_{3,2}}{n}} \times \mathbb{1}_{d_0 + \frac{h_{3,2}}{n} \le X_j < d_0 + \frac{h_{3,3}}{n}}\right] \\
&\le C_{k,\theta_0} \times n^2 \times \mathbb{P}\left(d_0 + \frac{h_{3,1}}{n} \le X < d_0 + \frac{h_{3,2}}{n}\right) \\
&\qquad\qquad \times \mathbb{P}\left(d_0 + \frac{h_{3,2}}{n} \le X < d_0 + \frac{h_{3,3}}{n}\right) \\
&\le C_{k,\theta_0}C^2 \times (h_{3,2} - h_{3,1}) \times (h_{3,3} - h_{3,2}) \qquad [C = \max_x f_X(x)] \\
&\le C_{k,\theta_0}C^2 \times (h_{3,3} - h_{3,1})^2
\end{aligned}
$$

This completes the proof of tightness of $\tilde{\tilde{\mathbb{Q}}}_n$. Hence using equation (A.33) we conclude:

$$\tilde{\tilde{\mathbb{Q}}}_n|_{\mathbb{I}} \overset{\mathscr{L}}{\Longrightarrow} \mathbb{Q}|_{\mathbb{I}}\,.$$

which, along with equation (A.32) implies:

$$\tilde{\mathbb{Q}}_n|_{\mathbb{I}} \overset{\mathscr{L}}{\Longrightarrow} \mathbb{Q}|_{\mathbb{I}}\,. \tag{A.34}$$

Finally, from the decomposition in equation (A.18) and combining our conclusions from equation (A.19), (A.26) and equation (A.34) we have:

$$\mathbb{Q}_n|_{\mathbb{I}} \overset{\mathscr{L}}{\Longrightarrow} \mathbb{Q}|_{\mathbb{I}}\,. \tag{A.35}$$

Next and last step is to invoke argmin continuity mapping theorem to say that:

$$\widehat{\mathbf{h}}_n = \left(\sqrt{n}\left(\hat{\alpha} - \alpha_0\right), \sqrt{n}\left(\hat{\beta} - \beta_0\right), n\left(\hat{d} - d_0\right)\right) \overset{\mathscr{L}}{\Longrightarrow} \text{mid argmin}_{\mathbf{h} \in \mathbb{R}^3}\mathbb{Q}(\mathbf{h})\,.$$

This will complete the proof. Following the proof of Lemma 3.2 of [26], all we need to establish the joint asymptotic tightness of $\{(\mathbb{Q}_n(\mathbf{h}), \mathbb{J}_n(\mathbf{h}))\}_{n\in\mathbb{N}}$ where $\mathbb{J}_n(\mathbf{h})$ is the jump process corresponding to $\mathbb{Q}_n(\mathbf{h})$, i.e.

$$\mathbb{J}_n(\mathbf{h}) = \text{sign}(h_3) \sum_{i=1}^{n} \left[\mathbb{1}_{X_i \le d_0 + \frac{h_3}{n}} - \mathbb{1}_{X_i \le d_0} \right] .$$

As we have already established the tightness of $\{\mathbb{Q}_n(\mathbf{h})\}$, we only need to establish the tightness of $\{\mathbb{J}_n(\mathbf{h})\}$. The proof is very similar to the proof of Lemma 3.2 of [26] and skipped here for the brevity.

A.6. Proofs of Theorem 2.2 and 2.3

The proofs of Theorem 2.2 and 2.3 are similar to that of Theorem 2.1 by replacing $\tilde{H}_k$ with ℓ_1 loss function and ℓ_2 loss function respectively and hence skipped.

A.7. Proof of Theorem 3.7

Here we assume $\alpha_0 = 0, \beta_0 = 1$ is known and derive the bound on the $\hat{d}$. When α_0, β_0 is not known, then the problem becomes harder and rate of convergence obviously can not be faster. Our proof is based on the proof of Theorem 5 of [30]. Consider $\mathscr{A}$ to be the set of all half-spaces, i.e.

$$\mathscr{A} = \left\{ x^\top d > 0 \right\}_{d \in S^{p-1}} .$$

Our model is $X \sim P$ and:

$$Y = \mathbb{1}_{X^\top d > 0} + \xi$$

where $\xi \sim \mathcal{N}(0,1)$. Now the class of hyperplanes $\mathscr{A}$ has VC dimension p. From the properties of the hyperplane we know that, given any $N > p$ (not to be confused with sample size n) there exist $x_1, \ldots, x_N \in \mathbb{R}^p$ such that $\mathscr{A}$ shatters all subsets of $\{x_1, \ldots, x_N\}$ with cardinality $k \le \lfloor p/2 \rfloor := V$ (e.g. see [13]). Define $\Theta_{N,V}$ to be the collection of all such $d \in S^{p-1}$ which shatters all subsets of length V of $\{x_1, \ldots, x_N\}$. Hence we have:

$$\left\{ \left(\mathbb{1}_{x_1^\top d > 0}, \ldots, \mathbb{1}_{x_N^\top d > 0} \right) \right\}_{d \in \Theta_{N,V}} := B$$

where:

$$B = \{0,1\}_{N,V} = \left\{ b \in \{0,1\}^N : \sum_{i=1}^{N} b_i = V \right\} .$$

Henceforth for any $d \in \Theta_{N,V}$ we denote by b_d to be the corresponding unique $b \in B$. Define μ to be the uniform measure on $\{x_1, \ldots, x_N\}$ and for any $d \in \Theta_{N,V}$ define:

$$Y = \mathbb{1}_{X^\top d > 0} + \xi \, .$$

The loss function we use here is the squared error loss defined as:

$$\begin{aligned}\ell(d, d_0) &= \mathbb{E}\left[(Y - \mathbb{1}_{X^\top d>0})^2 - (Y - \mathbb{1}_{X^\top d_0>0})^2\right] \\ &= \mathbb{E}_X\left[|\mathbb{1}_{X^\top d>0} - \mathbb{1}_{X^\top d_0>0}|\right] \\ &= \left\|\mathbb{1}_{X^\top d>0} - \mathbb{1}_{X^\top d_0>0}\right\|_{L_1(P)} .\end{aligned}$$

The minimax risk is defined as:

$$\begin{aligned}\mathcal{R}_n &= \inf_{\hat{d}} \sup_{d\in S^{p-1}} \mathbb{E}_d\left[\ell(\hat{d}, d)\right] \\ &\geq \inf_{\hat{d}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\ell(\hat{d}, d)\right] \\ &= \inf_{\hat{d}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\left\|\mathbb{1}_{X^\top \hat{d}>0} - \mathbb{1}_{X^\top d>0}\right\|_{L_1(\mu)}\right] \\ &= \frac{1}{N}\inf_{\hat{d}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\sum_{i=1}^N \left|\mathbb{1}_{x_i^\top \hat{d}>0} - \mathbb{1}_{x_i^\top d>0}\right|_{L_1(\mu)}\right] \qquad \text{(A.36)}\end{aligned}$$

Now for any $\hat{d} \in S^{p-1}$, define $\hat{d}_{\text{new}}$ as:

$$\hat{d}_{\text{new}} = \operatorname*{argmin}_{d_*\in\Theta_{N,V}} \left\|\mathbb{1}_{X^\top \hat{d}>0} - \mathbb{1}_{X^\top d_*>0}\right\|_{L_1(\mu)} .$$

Then we have for any $d \in \Theta_{N,V}$:

$$\begin{aligned}\left\|\mathbb{1}_{X^\top \hat{d}_{\text{new}}>0} - \mathbb{1}_{X^\top d>0}\right\|_{L_1(\mu)} &= \left\|\mathbb{1}_{X^\top \hat{d}_{\text{new}}>0} - \mathbb{1}_{X^\top \hat{d}>0} + \mathbb{1}_{X^\top \hat{d}>0} - \mathbb{1}_{X^\top d>0}\right\|_{L_1(\mu)} \\ &\leq \left\|\mathbb{1}_{X^\top \hat{d}_{\text{new}}>0} - \mathbb{1}_{X^\top \hat{d}>0}\right\|_{L_1(\mu)} \\ &\quad + \left\|\mathbb{1}_{X^\top \hat{d}>0} - \mathbb{1}_{X^\top d>0}\right\|_{L_1(\mu)} \\ &\leq 2\left\|\mathbb{1}_{X^\top \hat{d}>0} - \mathbb{1}_{X^\top d>0}\right\|_{L_1(\mu)} .\end{aligned}$$

Putting this bound in equation (A.36) we obtain:

$$\begin{aligned}\mathcal{R}_n &\geq \frac{1}{N}\inf_{\hat{d}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\sum_{i=1}^N \left|\mathbb{1}_{x_i^\top \hat{d}>0} - \mathbb{1}_{x_i^\top d>0}\right|_{L_1(\mu)}\right] \\ &\geq \frac{1}{2N}\inf_{\hat{d}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\sum_{i=1}^N \left|\mathbb{1}_{x_i^\top \hat{d}_{\text{new}}>0} - \mathbb{1}_{x_i^\top d>0}\right|_{L_1(\mu)}\right] \\ &= \frac{1}{2N}\inf_{\hat{d}\in\Theta_{N,V}} \sup_{d\in\Theta_{N,V}} \mathbb{E}_d\left[\sum_{i=1}^N \left|\mathbb{1}_{x_i^\top \hat{d}>0} - \mathbb{1}_{x_i^\top d>0}\right|_{L_1(\mu)}\right] \qquad \text{(A.37)}\end{aligned}$$

Next note that:

$$\sum_{i=1}^N \left|\mathbb{1}_{x_i^\top \hat{d}>0} - \mathbb{1}_{x_i^\top d>0}\right| = d_H\left(b_{\hat{d}}, b_d\right)$$

where d_H is the Hamming distance. As $\Theta_{N,V}$ has a bijection with B we can write equation (A.37) as:

$$\mathcal{R}_n \geq \frac{1}{2N} \inf_{\hat{b}} \sup_{b \in B} \mathbb{E}_b \left[d_H \left(\hat{b}, b \right) \right] \geq \frac{1}{2N} \inf_{\hat{b} \in \mathcal{D}} \sup_{b \in \mathcal{D}} \mathbb{E}_b \left[d_H \left(\hat{b}, b \right) \right] \tag{A.38}$$

for any subset $\mathcal{D} \subseteq B$. We now choose $\mathcal{D}$ carefully. Note that for any $N \geq 4V$ (i.e. $N \geq 2p$), we can choose $\mathcal{D}$ such that (Lemma 8 of [34]):

1. $d_H(b, b') > \frac{V}{2}$ for all $b \neq b' \in \mathcal{D}$.
2. $\log |\mathcal{D}| \geq \rho V \log \left(\frac{N}{V} \right)$ with $\rho = 0.233$.

Using this we modify equation (A.38) as follows:

$$\begin{aligned}
\mathcal{R}_n &\geq \frac{1}{2N} \inf_{\hat{b} \in \mathcal{D}} \sup_{b \in \mathcal{D}} \mathbb{E}_b \left[d_H \left(\hat{b}, b \right) \right] \\
&= \frac{1}{2N} \inf_{\hat{b} \in \mathcal{D}} \sup_{b \in \mathcal{D}} \mathbb{E}_b \left[d_H \left(\hat{b}, b \right) \mathbb{1}_{\hat{b} \neq b} \right] \\
&\geq \frac{V}{4N} \inf_{\hat{b} \in \mathcal{D}} \sup_{b \in \mathcal{D}} \mathbb{P}_b \left(\hat{b} \neq b \right) \qquad \text{[From point 1. above]} \\
&= \frac{V}{4N} \inf_{\hat{b} \in \mathcal{D}} \sup_{b \in \mathcal{D}} \left(1 - \mathbb{P}_b \left(\hat{b} = b \right) \right) \\
&= \frac{V}{4N} \inf_{\hat{b} \in \mathcal{D}} \left(1 - \min_{b \in \mathcal{D}} \mathbb{P}_b \left(\hat{b} = b \right) \right)
\end{aligned} \tag{A.39}$$

Now to further bound the above equation, we use the following lemma (see [8]):

Lemma A.9. *Let $m \geq 1$, $(P_i)_{0 \leq i \leq m}$ be a family of probability distributions and $(A_i)_{0 \leq i \leq m}$ be a family of disjoint events. Let $a = \min_{0 \leq i \leq m} P_i(A_i)$. Then setting:*

$$\bar{\mathscr{K}} = \frac{1}{m} \sum_{i=1}^{m} \mathscr{K}(P_i, P_0)$$

where $\mathscr{K}$ is the Kullback-Liebler divergence, we have:

$$a \leq \alpha \vee \left(\frac{\bar{\mathscr{K}}}{\log(1+m)} \right).$$

where $\alpha = 0.71$.

We now use this bound in equation (A.39). Fix any $b_0 \in \mathcal{D}$. Define $A_i = \left\{ \hat{b} = b_i \right\}$ for all $b_i \in \mathcal{D}$ which are disjoint events. Hence using Lemma A.9 we obtain:

$$\min_{b \in \mathcal{D}} \mathbb{P}_b \left(\hat{b} = b \right) \leq \alpha \vee \left(\frac{\bar{\mathscr{K}}}{\log |\mathcal{D}|} \right) \tag{A.40}$$

where:

$$\bar{\mathscr{K}} = \frac{1}{|\mathcal{D}| - 1} \sum_{b \in \mathcal{D}, b \neq b_0} \mathscr{K} \left(P_b^{\otimes n}, P_{b_0}^{\otimes n} \right) = \frac{n}{|\mathcal{D}| - 1} \sum_{b \in \mathcal{D}, b \neq b_0} \mathscr{K} \left(P_b, P_{b_0} \right).$$

Now note that for any d_1, d_2 we have:

$$\begin{aligned}\mathscr{K}\left(P_{d_1}, P_{d_2}\right) &= \mathbb{E}_X\left[\mathscr{K}\left(P_{d_1}(Y \mid X), P_{d_2}(Y \mid X)\right)\right] \\ &= \mathbb{E}_X\left[\left(\mathbb{1}_{X^\top d_1 > 0} - \mathbb{1}_{X^\top d_2 > 0}\right)^2\right] \\ &= \mathbb{E}_X\left[\left|\mathbb{1}_{X^\top d_1 > 0} - \mathbb{1}_{X^\top d_2 > 0}\right|\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{1}_{x_i^\top d_1 > 0} - \mathbb{1}_{x_i^\top d_2 > 0}\right| \\ &= \frac{1}{N} d_H(b_{d_1}, b_{d_2})\,.\end{aligned}$$

Also by definition for any two $b, b' \in \mathcal{D}$ we have $d_H(b, b') \leq 2V$. Hence we have:

$$\bar{\mathscr{K}} \leq \frac{2Vn}{N}\,.$$

This bound along with equation (A.40) modifies the bound of equation (A.39) as:

$$\begin{aligned}\mathcal{R}_n &\geq \frac{V}{4N}\inf_{\hat{b}\in\mathcal{D}}\left(1 - \min_{b\in\mathcal{D}}\mathbb{P}_b\left(\hat{b} = b\right)\right) \\ &\geq \frac{V}{4N}\left(1 - \left(\alpha \vee \frac{2Vn}{N\log|\mathcal{D}|}\right)\right) \\ &= \frac{V(1-\alpha)}{4N} \qquad \text{(A.41)}\end{aligned}$$

when,

$$\alpha \geq \frac{2Vn}{N\log|\mathcal{D}|}\,.$$

which holds if:

$$\alpha > \frac{2Vn}{N\rho V\log\left(N/V\right)}$$

i.e. if:

$$N\log\left(\frac{N}{V}\right) \geq \frac{2n}{\alpha\rho}\,. \qquad \text{(A.42)}$$

which is satisfied, if for example:

$$N = \left\lfloor \frac{4n}{\rho\alpha\left(1 + \log\left(\frac{n}{V}\right)\right)} \right\rfloor.$$

Using this in equation (A.41) we conclude:

$$\mathcal{R}_n \gtrsim \frac{V}{n}\left(1 + \log\left(\frac{n}{V}\right)\right) \asymp \frac{p}{n}\left(1 + \log\left(\frac{n}{p}\right)\right).$$

as $V = \lfloor p/2 \rfloor$. We finally need to verify $N > 4V$ and that it satisfies equation (A.42). The first one is obviously true for all large n as $n/p \to \infty$. For the second one, lets forget the $\lfloor \cdot \rfloor$ in the definition of N for the time being as it will not affect asymptotically. Then:

$$\begin{aligned}
& N \log\left(\frac{N}{V}\right) \geq \frac{2n}{\alpha\rho} \\
\iff & \frac{4n}{\rho\alpha\left(1+\log\left(\frac{n}{V}\right)\right)} \log\left(\frac{4n}{V\rho\alpha\left(1+\log\left(\frac{n}{V}\right)\right)}\right) \geq \frac{2n}{\alpha\rho} \\
\iff & \frac{1}{\left(1+\log\left(\frac{n}{V}\right)\right)} \log\left(\frac{4n}{V\rho\alpha\left(1+\log\left(\frac{n}{V}\right)\right)}\right) \geq \frac{1}{2} \\
\iff & \frac{1}{\left(1+\log\left(\frac{n}{V}\right)\right)} \left[\log\left(\frac{n}{V}\right) + \log\left(\frac{4}{\rho\alpha}\right) - \log\left(1+\log\left(\frac{n}{V}\right)\right)\right] \geq \frac{1}{2}
\end{aligned}$$

As $n/V \to \infty$, LHS converges to 1 and eventually $> 1/2$. Therefore the choice of N is valid for all large n.

A.8. Proof of Theorem 3.10

We first assume that Assumption 3.5 holds globally and our parameter space $\Omega = \Omega_\alpha \times \Omega_\beta$ for (α_0, β_0) is such that:

$$\min_{\alpha \in \Omega_\alpha} |\alpha - \beta_0| \wedge \min_{\beta \in \Omega_\beta} |\beta - \alpha_0| \geq \delta > 0\,.$$

This is just to avoid the issue of consistency. One can relax this assumption with an additionally showing that the estimators are consistent. To prove Theorem 3.10 we use Theorem A.3 of [31]. To match our notation with that theorem, here our loss function $\gamma(\theta, \cdot)$ is:

$$\begin{aligned}
\gamma(\theta, (X, \xi)) &= \tilde{H}_k\left(\xi + \alpha_0 \mathbb{1}_{X^\top d_0 \leq 0} + \beta_0 \mathbb{1}_{X^\top d_0 > 0} - \alpha \mathbb{1}_{X^\top d \leq 0} - \beta \mathbb{1}_{X^\top d > 0}\right) \\
&\qquad - \tilde{H}_k(\xi) \\
&= \left(\tilde{H}_k\left(\xi + \alpha_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \vee X^\top d_0 \leq 0} \\
&\quad \left(\tilde{H}_k\left(\xi + \alpha_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d_0 \leq 0 < X^\top d} \\
&\quad + \left(\tilde{H}_k\left(\xi + \beta_0 - \alpha\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \leq 0 < X^\top d_0} \\
&\quad + \left(\tilde{H}_k\left(\xi + \beta_0 - \beta\right) - \tilde{H}_k(\xi)\right) \mathbb{1}_{X^\top d \wedge X^\top d_0 > 0}
\end{aligned}$$

It is immediate from the definition that $\gamma(\theta_0, (X, \xi)) = 0$. Also from equation (A.17) we know $\gamma(\theta, \cdot)$ is uniform bounded by some constant only depending on k and the width of Ω. Following similar arguments used in the proof of Theorem 3.6, we obtain:

$$\ell(\theta, \theta_0) = \mathbb{E}[\gamma(\theta, (X, \xi))] \geq c_k \ \mathsf{dist}^2(\theta, \theta_0)\,, \tag{A.43}$$

for some constant c_k (independent of n), where the dist function is:

$$\mathsf{dist}(\theta,\theta_0)=\sqrt{(\alpha-\alpha_0)^2+(\beta-\beta_0)^2+\mathbb{P}\left(\mathrm{sign}(X^\top d)\neq\mathrm{sign}(X^\top d_0)\right)}\,.$$

Moreover, from equation (A.17) we have:

$$\mathsf{var}\left(\gamma(\theta,(X,\xi))\right)\leq\mathbb{E}[\gamma^2(\theta,(X,\xi))]\leq C_k\ \mathsf{dist}^2(\theta,\theta_0)\,.$$

Hence this semi-metric dist satisfies conditions of Theorem A.3 of [31] with $\omega(x)=x$. Next we need to bound the modulus of continuity:

$$\sqrt{n}\mathbb{E}\left[\sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\left|(\mathbb{P}_n-P)\left(\gamma(\theta,(X,\xi))-\gamma(\theta_m,(X,\xi))\right)\right|\right]$$

Note that another application of equation (A.17) yields:

$$\sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\mathbb{E}\left[\left(\gamma(\theta,(X,\xi))-\gamma(\theta_m,(X,\xi))\right)^2\right]\lesssim\epsilon^2\,.$$

Hence applying Theorem 8.7 of [35] we have:

$$\begin{aligned}&\sqrt{n}\mathbb{E}\left[\sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\left\|(\mathbb{P}_n-P)\left(\gamma(\theta,(X,\xi))-\gamma(\theta_m,(X,\xi))\right)\right\|\right] &\text{(A.44)}\\ &\lesssim\epsilon\sqrt{V_m\log\frac{1}{\epsilon}}\vee\frac{V_m}{\sqrt{n}}\log\frac{1}{\epsilon}\\ &:=\psi_m(\epsilon)\,. &\text{(A.45)}\end{aligned}$$

So a value of ϵ_m that satisfies $\sqrt{n}\epsilon_m^2\geq\phi_m(\epsilon_m)$ is:

$$\epsilon_m=\frac{V_m}{n}\log\frac{n}{V_m}\,.$$

Therefore we can take $x_m=V_m\log(n/V_m)$ and as $\omega(x)=x$ we have $b(n)=1$ (see Theorem A.3 of [31] for the exact expression of $\omega(x),b(n)$). Note that, as we are assuming $(s\log p)/n\to 0$ (Assumption 3.9), it is sufficient to search among the models with $1\leq m\leq C\lfloor(n/\log p)\rfloor$ for any constant C. We take $C=1/4$ here. With these choices, the value of Σ (as defined in Theorem A.3 of [31]) is:

$$\begin{aligned}\Sigma&=\sum_{i=1}^{\frac{1}{4}\lfloor\frac{n}{\log p}\rfloor}e^{-V_i\log\frac{n}{V_i}}\\ &\leq\lim_{n\to\infty}\sum_{i=1}^{\frac{1}{4}\lfloor\frac{n}{\log p}\rfloor}e^{-V_i\log\frac{n}{V_i}}<\infty\,.\end{aligned}$$

Hence, an application Theorem A.3 of [31] yields:

$$\mathbb{P}\left(\ell(\hat{\theta},\theta_0) > C\text{pen}(s) + t\frac{C_1}{n}\right) \leq \Sigma e^{-t}\,.$$

This along with the value of pen(s) from equation (3.4) of the main paper and the lower bound equation (A.43) completes the first part of the proof, i.e.

$$\mathsf{dist}^2\left((\hat{\alpha}_{\text{init}},\hat{\beta}_{\text{init}},\hat{d}),(\alpha_0,\beta_0,d_0)\right) = O_p\left(\frac{V_s}{n}\log\frac{n}{V_s}\right)\,.$$

The acceleration of the rate of $\hat{\alpha},\hat{\beta}$ via replacing d_0 by $\hat{d}$ in the model equation is exactly same as that of Theorem 3.6 and hence skipped.

A.9. Proof of Theorem 3.14

Here again we assume for technical simplicity that the wedge assumption (Assumption 3.5) is valid on entire S^{p-1}, although all our arguments can be extended to the case where the assumption is valid only locally along with a separate argument for the consistency of the estimator. We use the same notations as of Theorem 3.10 through out the proof. Define

$$y_m^2 = 2\kappa\frac{V_m\|\xi\|_{n,2}}{n}\log\frac{n}{V_m}$$

for all $m\in\mathcal{M}$, where V_m is the VC dimension of model m. For any such model m we obtain $\hat{\theta}_m$ as:

$$\begin{aligned}
\hat{\theta}_m &= \underset{\theta\in\Omega\times S_m^{p-1}}{\text{argmin}}\ \mathbb{P}_n\left\{(Y-\alpha\mathbb{1}_{X^\top d\leq 0}-\beta\mathbb{1}_{X^\top d>0})^2-\xi^2\right\}\\
&= \underset{\theta\in\Omega\times S_m^{p-1}}{\text{argmin}}\ \mathbb{P}_n\Big\{\xi\left(\alpha_0\mathbb{1}_{X^\top d_0\leq 0}+\beta_0\mathbb{1}_{X^\top d_0>0}-\alpha\mathbb{1}_{X^\top d\leq 0}-\beta\mathbb{1}_{X^\top d>0}\right)\\
&\qquad +\frac{1}{2}\left(\alpha_0\mathbb{1}_{X^\top d_0\leq 0}+\beta_0\mathbb{1}_{X^\top d_0>0}-\alpha\mathbb{1}_{X^\top d\leq 0}-\beta\mathbb{1}_{X^\top d>0}\right)^2\Big\}\\
&:= \underset{\theta\in\Omega\times S_m^{p-1}}{\text{argmin}}\ \mathbb{P}_n f_\theta\\
&:= \underset{\theta\in\Omega\times S_m^{p-1}}{\text{argmin}}\ \mathbb{P}_n\left(f_{\theta,1}+f_{\theta,2}\right)
\end{aligned}$$

where the functions $f_{d,1}, f_{d,2}$ are defined as:

$$\begin{aligned}
f_{\theta,1} &= \xi\left(\alpha_0\mathbb{1}_{X^\top d_0\leq 0}+\beta_0\mathbb{1}_{X^\top d_0>0}-\alpha\mathbb{1}_{X^\top d\leq 0}-\beta\mathbb{1}_{X^\top d>0}\right)\,,\\
f_{\theta,2} &= \frac{1}{2}\left(\alpha_0\mathbb{1}_{X^\top d_0\leq 0}+\beta_0\mathbb{1}_{X^\top d_0>0}-\alpha\mathbb{1}_{X^\top d\leq 0}-\beta\mathbb{1}_{X^\top d>0}\right)^2\,.
\end{aligned}$$

The loss function used here is:

$$\ell(\theta,\theta_0) = \mathbb{P}f_\theta \gtrsim \mathsf{dist}^2(\theta,\theta_0)\,,$$

for all $\theta \in \Omega \in S^{p-1}$ via the global version of Assumption 3.5. From the definition of $\hat{m}$ we have:

$$\begin{aligned}\mathbb{P}_n f_{\hat{\theta}_{\hat{m}}} + \text{pen}(\hat{m}) &\leq \mathbb{P}_n f_{\hat{\theta}_{s_0}} + \text{pen}(s_0) \\ &\leq \mathbb{P}_n f_{\theta_{s_0}} + \text{pen}(s_0) := \text{pen}(s_0)\,.\end{aligned}$$

Using this we can bound the loss function:

$$\begin{aligned}\ell(\hat{\theta}_{\hat{m}}, \theta_0) &= \mathbb{P} f_{\hat{\theta}_{\hat{m}}} \\ &= (\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_{\hat{m}}} + \mathbb{P}_n f_{\hat{\theta}_{\hat{m}}} \\ &= (\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_{\hat{m}}} + \mathbb{P}_n f_{\hat{\theta}_{\hat{m}}} + \text{pen}(\hat{m}) - \text{pen}(\hat{m}) \\ &\leq (\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_{\hat{m}}} + \text{pen}(s_0) - \text{pen}(\hat{m}) \\ &= \frac{(\mathbb{P} - \mathbb{P}_n) f_{\hat{\theta}_{\hat{m}}}}{\ell(\hat{\theta}_{\hat{m}}, \theta_0) + y_{\hat{m}}^2} \left(\ell(\hat{\theta}_{\hat{m}}, \theta_0) + y_{\hat{m}}^2\right) + \text{pen}(s_0) - \text{pen}(\hat{m}) \\ &\leq \sup_{\theta \in \Omega \times S_{\hat{m}}^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n) f_\theta|}{\ell(\theta, \theta_0) + y_{\hat{m}}^2} \left(\ell(\hat{\theta}_{\hat{m}}, d_0) + y_{\hat{m}}^2\right) + \text{pen}(s_0) - \text{pen}(\hat{m})\end{aligned}$$

For the rest of the calculation, define:

$$\begin{aligned}\Gamma_m &= \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n) f_\theta|}{\ell(\theta, \theta_0) + y_m^2} \\ &= \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n)(f_{\theta,1} + f_{\theta,2})|}{\ell(\theta, \theta_0) + y_m^2} \\ &\leq \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n) f_{\theta,1}|}{\ell(\theta, \theta_0) + y_m^2} + \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n) f_{\theta,2}|}{\ell(\theta, \theta_0) + y_m^2} \\ &:= \Gamma_{m,1} + \Gamma_{m,2}\,.\end{aligned}$$

Next we try to bound $\Gamma_{\hat{m}}$. More specifically, we bound Γ_m for all m and then use a union bound to bound $\Gamma_{\hat{m}}$. Note that, as the function class under the consideration of $\Gamma_{m,2}$ is bounded we can use similar as of the proof of Theorem A.3 of [31] (i.e. applying Talagrand's inequality and then bound the expectation and variance) to conclude:

$$\mathbb{P}\left(\Gamma_{\hat{m},2} \geq 1/4\right) = o(1)\,. \tag{A.46}$$

For $\Gamma_{m,1}$, we first decompose it as follows:

$$\begin{aligned}\Gamma_{m,1} &\leq \sup_{d \in S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n)(f_{\theta,1} - f_{\theta_m,1})|}{\ell(\theta, \theta_0) + y_m^2} + \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n) f_{\theta_m,1}|}{\ell(\theta, \theta_0) + y_m^2} \\ &\leq \sup_{\theta \in \Omega \times S_m^{p-1}} \frac{|(\mathbb{P} - \mathbb{P}_n)(f_{\theta,1} - f_{\theta_m,1})|}{\ell(\theta, \theta_0) + y_m^2} + \frac{|(\mathbb{P} - \mathbb{P}_n) f_{\theta_m,1}|}{\ell(\theta_m, \theta_0) + y_m^2} \\ &= \Gamma_{m,11} + \Gamma_{m,12}\,.\end{aligned}$$

Bounding $\mathbb{E}[\Gamma_{m,12}]$ is straight-forward:

$$\begin{aligned}\mathbb{E}[\Gamma_{m,12}] &\leq \frac{\sqrt{\mathsf{var}\left(f_{\theta_m,1}\right)}}{\sqrt{n}\left(\ell(\theta_m,\theta_0)+y_m^2\right)} \\ &\leq \frac{\|\xi\|_2 \mathsf{dist}(\theta_m,\theta_0)}{\sqrt{n}\left(\mathsf{dist}^2(\theta_m,\theta_0)+y_m^2\right)} \\ &\leq \frac{\|\xi\|_2}{\sqrt{n}} \sup_{x\geq 0}\frac{x}{x^2+y_m^2} \leq \frac{\|\xi\|_2}{2}\frac{1}{\sqrt{n}y_m}\,.\end{aligned}$$

To bound $\mathbb{E}[\Gamma_{m,1}]$ we use the maximal inequality of the weighted empirical process (see Lemma A.5 of [30]), which is a variant of peeling argument. First of all note that, by symmetrization and applying Theorem 8.7 of [35], we have for any $1 \leq k \leq n$:

$$\begin{aligned}&\mathbb{E}\left[\sup_{\substack{\theta\in\Omega\times S_m^{p-1}\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\left|\sum_{i=1}^{k}\xi_i\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0}+\beta_m\mathbb{1}_{X_i^\top d_m>0}-\alpha\mathbb{1}_{X_i^\top d\leq 0}-\beta\mathbb{1}_{X_i^\top d>0}\right)\right|\right] \\ &\qquad\lesssim \left(\sigma_\epsilon\sqrt{kV_m\log\frac{1}{\sigma_\epsilon}}\vee V_m\log\frac{1}{\sigma_\epsilon}\right)\,.\end{aligned}$$

where the *wimpy variance* σ_ϵ^2 is defined as:

$$\begin{aligned}\sigma_\epsilon^2 &= \sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\sigma_\xi^2\,\mathbb{E}\left[\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0}+\beta_m\mathbb{1}_{X_i^\top d_m>0}-\alpha\mathbb{1}_{X_i^\top d\leq 0}-\beta\mathbb{1}_{X_i^\top d>0}\right)^2\right] \\ &\lesssim \epsilon^2\,.\end{aligned}$$

This, along with Proposition A.7 implies:

$$\begin{aligned}&\mathbb{E}\left[\sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\left|\left(\mathbb{P}-\mathbb{P}_n\right)\left(f_{\theta,1}-f_{\theta_m,1}\right)\right|\right] \\ &= \mathbb{E}\left[\sup_{\substack{\theta:f_\theta\in\mathcal{F}_m\\ \mathsf{dist}(\theta,\theta_m)\leq\epsilon}}\left|\sum_{i=1}^{k}\xi_i\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0}+\beta_m\mathbb{1}_{X_i^\top d_m>0}\right.\right.\right. \\ &\qquad\qquad\left.\left.\left.-\alpha\mathbb{1}_{X_i^\top d\leq 0}-\beta\mathbb{1}_{X_i^\top d>0}\right)\right|\right] \\ &\lesssim \|\xi\|_{2,1}\,\epsilon\sqrt{\frac{V_m}{n}\log\left(\frac{1}{\epsilon}\right)}+2\frac{V_m}{n}\log\left(\frac{1}{\epsilon}\right)\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] \\ &:= \frac{\psi_m(\epsilon)}{\sqrt{n}}\,.\end{aligned}$$

An application of Lemma A.5 of [30] yields:

$$\mathbb{E}[\Gamma_{m,1}] \lesssim \frac{\psi_m(2\sqrt{2}y_m)}{\sqrt{n}y_m^2}$$

which, in turn, yields:

$$\begin{aligned}
E(\Gamma_{m,1}) &\leq \frac{4\phi_m(2\sqrt{2}y_m)}{\sqrt{n}y_m^2} + \frac{\|\xi\|_2}{2\sqrt{n}y_m} \\
&\leq \frac{4\phi_m(2\sqrt{2}\epsilon_m)}{\sqrt{n}y_m\epsilon_m} + \frac{\|\xi\|_2}{2\sqrt{n}y_m}\sqrt{\frac{y_m^2}{y_m^2}} \\
&\leq \frac{8\sqrt{2}\phi_m(\epsilon_m)}{\sqrt{n}y_m\epsilon_m} + \frac{\|\xi\|_2}{2\sqrt{n}y_m}\sqrt{\frac{\phi_m^2(y_m)}{y_m^2}} \\
&\leq \frac{8\sqrt{2}\epsilon_m}{y_m} + \frac{2}{\sqrt{n}y_m}\sqrt{\frac{\phi_m^2(\epsilon_m)}{\epsilon_m^2}} \\
&\leq \frac{8}{\sqrt{\kappa}} + \frac{\|\xi\|_2}{2\sqrt{n}y_m}\frac{\phi_m(\epsilon_m)}{\epsilon_m} \\
&\leq \frac{8}{\sqrt{\kappa}} + \frac{\|\xi\|_2\epsilon_m}{2y_m} \leq \frac{8+\sqrt{2}}{\sqrt{\kappa}}
\end{aligned} \tag{A.47}$$

which can be made arbitrarily small by making κ arbitrarily large. Next, we bound the fluctuation of $V_{m,1}$ around its mean using Chebychev inequality:

$$\mathbb{P}\left(|\Gamma_{m,1} - \mathbb{E}[\Gamma_{m,1}]| \geq t\right) \leq \frac{\mathsf{var}(\Gamma_{m,1})}{t^2}.$$

To bound the variance we use Theorem 11.17 along with Theorem 11.1 of [9]. To match with their notation for the ease of the readers, we have:

$$\begin{aligned}
X_{i,\theta} = \frac{1}{n}\Bigg[&\frac{\xi_i\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0} + \beta_m\mathbb{1}_{X_i^\top d_m>0} - \alpha\mathbb{1}_{X_i^\top d\leq 0} - \beta\mathbb{1}_{X_i^\top d>0}\right)}{\ell(\theta,\theta_0)+y_m^2} \\
&-\mathbb{E}\left(\frac{\xi_i\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0} + \beta_m\mathbb{1}_{X_i^\top d_m>0} - \alpha\mathbb{1}_{X_i^\top d\leq 0} - \beta\mathbb{1}_{X_i^\top d>0}\right)}{\ell(\theta,\theta_0)+y_m^2}\right)\Bigg] \\
= \frac{1}{n}\Bigg[&\frac{\xi_i\left(\alpha_m\mathbb{1}_{X_i^\top d_m\leq 0} + \beta_m\mathbb{1}_{X_i^\top d_m>0} - \alpha\mathbb{1}_{X_i^\top d\leq 0} - \beta\mathbb{1}_{X_i^\top d>0}\right)}{\ell(\theta,\theta_0)+y_m^2}\Bigg]
\end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}(\xi) = 0$. That $X_{i,\theta}$ is symmetric follows from the symmetry of ξ. We define M as:

$$\max_{1\leq i\leq n}\sup_{\theta\in\Omega\times S_m^{p-1}} X_{i,\theta}^2 \lesssim \frac{\max_{1\leq i\leq n}\xi_i^2}{n^2y_m^4} := M$$

and the wimpy variance:

$$\begin{aligned}\sup_{\theta\in\Omega\times S_m^{p-1}}\sum_{i=1}^{n}\mathbb{E}(X_{i,\theta}^2) &\le \frac{2\sigma_\xi^2}{n}\frac{\ell(\theta_m,\theta_0)}{(\ell(\theta_m,\theta_0)+y_m^2)^2}\\ &\le \frac{2\sigma_\xi^2}{n}\sup_{x\ge 0}\frac{x}{(x+y_m^2)^2}\\ &\le \frac{2\sigma_\xi^2}{4ny_m^2} := \sigma_m^2\,.\end{aligned}$$

An application of Theorem 11.17 and Theorem 11.1 of [9] yields:

$$\mathsf{var}(\Gamma_{m,1}) \le \sigma_m^2 + 64\sqrt{\mathbb{E}[M_m]}\mathbb{E}[\Gamma_{m,1}] + 18^2\mathbb{E}[M_m]\,. \tag{A.48}$$

Note that we set $V_m = m(\log p)^{1+\delta}$ (which is slightly larger than the VC dimension) and choose y_m^2 as:

$$y_m^2 = 2\frac{V_m\|\xi\|_{2,n}}{n}\log\frac{n}{V_m} = 2\mathrm{pen}(m)\,.$$

As per Assumption 3.13, we confine the model selection in

$$1 \le m \le (1/4)\lfloor n/(\log p)^2\rfloor\,.$$

To facilitate the union, we next show that $\sum_{i=1}^{\mathcal{M}}\mathsf{var}(V_{m,1}) \to 0$ as $n\to\infty$. We bound each terms on RHS of equation (A.48):

$$\begin{aligned}\sum_{i=1}^{\mathcal{M}}\sigma_m^2 &= \frac{\sigma_\xi^2}{4}\sum_{i=1}^{\mathcal{M}}\frac{1}{ny_m^2}\\ &= \frac{\sigma_\xi^2}{4\|\xi\|_{n,2}}\sum_{i=1}^{n/(4(\log p)^2)}\frac{1}{V_m\log\frac{n}{V_m}}\\ &\le \frac{\sigma_\xi^2}{4\log 4\|\xi\|_{n,2}}\sum_{i=1}^{n/4(\log p)^2}\frac{1}{V_m}\\ &\le \frac{\sigma_\xi^2}{4\log 4(\log p)^{\delta/2}\|\xi\|_{n,2}}\sum_{i=1}^{n/(4(\log p)^2)}\frac{1}{m(\log m)^{1+\delta/2}}\\ &\le \frac{\sigma_\xi^2}{4\log 4(\log p)^{\delta/2}\|\xi\|_{n,2}}\sum_{i=1}^{\infty}\frac{1}{m(\log m)^{1+\delta/2}}\\ &\longrightarrow 0 \text{ as } n\to\infty.\end{aligned}$$

Now for the second summand:

$$64\sum_{i=1}^{\mathcal{M}}\sqrt{\mathbb{E}[M_m]}\mathbb{E}[\Gamma_{m,1}] \le 16\sqrt{\mathbb{E}[M_m]}$$

$$\begin{aligned}
&\lesssim 8\sum_{i=1}^{\mathcal{M}}\sqrt{\frac{\mathbb{E}[\max_{1\le i\le n}\xi_i^2]}{n^2y_m^4}}\\
&= 8\sum_{i=1}^{\mathcal{M}}\frac{\|\xi\|_{n,2}}{ny_m^2}\\
&\le \frac{8}{4\log 4(\log p)^{\delta/2}}\sum_{i=1}^{\infty}\frac{1}{m(\log m)^{1+\delta/2}}\\
&\longrightarrow 0 \text{ as } n\to\infty.
\end{aligned}$$

And similarly for the third summand:

$$\begin{aligned}
18^2\sum_{i=1}^{\mathcal{M}}\mathbb{E}[M_m] &= 18^2\sum_{i=1}^{\mathcal{M}}\frac{\mathbb{E}[\max_{1\le i\le n}\xi_i^2]}{n^2y_m^4}\\
&\le \frac{18^2}{4\log 4(\log p)^{\delta}}\sum_{i=1}^{\infty}\frac{1}{m^2(\log m)^{2+\delta}}\\
&\longrightarrow 0 \text{ as } n\to\infty.
\end{aligned}$$

Hence taking $t = 1/8$ and using the fact that $\mathbb{E}[\Gamma_{m,1}] \le 1/8$ for all m for our choice of κ, we have:

$$\mathbb{P}\left(\Gamma_{\hat{m},1} > \frac{1}{4}\right) \longrightarrow 0\,. \tag{A.49}$$

Therefore, combining equation (A.46) and (A.49) we conclude:

$$\mathbb{P}\left(\Gamma_{\hat{m}} > \frac{1}{2}\right) \longrightarrow 0$$

Hence, on its complement event, we have

$$\begin{aligned}
\ell(\hat{\theta}_{\hat{m}},\theta_0) &\le \frac{1}{2}\left(\ell(\hat{\theta}_{\hat{m}},d_0) + y_{\hat{m}}^2\right) + \text{pen}(s_0) - \text{pen}(\hat{m})\\
&= \frac{1}{2}\ell(\hat{\theta}_{\hat{m}},d_0) + \text{pen}(s_0)\,,
\end{aligned}$$

which further implies,

$$\ell(\hat{\theta}_{\hat{m}},\theta_0) \le 2\text{pen}(s_0)\,.$$

This, along with equation (A.43) indicates:

$$\mathsf{dist}^2\left((\hat{\alpha}_{\text{init}},\hat{\beta}_{\text{init}},\hat{d}),(\alpha_0,\beta_0,d_0)\right) = O_p\left(\frac{s_0(\log p)^{(1+\delta)}\|\xi\|_{n,2}}{n}\left(\log\frac{n}{s_0\log p}\right)\right)\,.$$

The boosting of the rate of $\hat{\alpha},\hat{\beta}$ by replacing d_0 by $\hat{d}$ in the model equation is exactly same as that of Theorem 3.6 and hence skipped.

A.10. Proof of Theorem 3.12

For the proof of this theorem we follow the techniques of proof of Theorem 2.18 of [31]. Recall Fano's inequality: if $\Theta \subseteq S^{p-1}$ is a finite 2ϵ packing set, i.e. for any two $d_i, d_j \in \Theta$, we have $\|d_i - d_j\| \geq 2\epsilon$ with $|\Theta| < \infty$, then based on n i.i.d. observations $z_1, \ldots, z_n$ we have the following minimax lower bound in estimating d_0:

$$\inf_{\hat{d}} \sup_{\mathcal{P}_d} \mathbb{E}\left[\left\|\hat{d} - d\right\|^2\right] \geq \epsilon^2 \left(1 - \frac{\frac{n}{M^2}\sum_{i,j:d_i,d_j\in\Theta} KL(P_{d_i}||P_{d_j}) + \log 2}{\log\left(|\Theta| - 1\right)}\right)$$

Next recall Gilbert-Varshamov Lemma: if d_H is the Hamming distance, i.e. $d_H(x,y) = \sum_{i=1}^{d} \mathbb{1}(x_i \neq y_i)$ with d being the ambient dimension. Then given any v with $1 \leq v \leq p/8$, we can find $\omega_1, \ldots, \omega_M \in \{0,1\}^p$ which satisfy the following:

1. $d_H(\omega_i, \omega_j) \geq \frac{v}{2} \quad \forall\, i \neq j \quad \in \{1, \ldots, m\}$.
2. $\log M \geq \frac{v}{8} \log\left(1 + \frac{d}{2v}\right)$.
3. $\|\omega_j\|_0 = v \quad \forall\, j \ \in \{1, \ldots, M\}$.

We choose the appropriate ϵ later. First, for a fixed $0 < \epsilon < 1$, we construct the set Θ as follows: applying Gilbert-Varshamov Lemma in dimension $p-1$ with sparsity $v = s-1$, we choose $\Omega = \{\omega_1, \ldots, \omega_M\} \in \{0,1\}^{p-1}$ which satisfies the above conditions (a) - (c). Then, for each $\omega_j \in \Omega$ set d_j as:

$$d_j = \frac{\left(1, \frac{\epsilon}{\sqrt{s-1}}\omega_j\right)}{\sqrt{1+\epsilon^2}}$$

It is immediate that $\|d_j\|_2 = 1$ and $\|d_j\|_0 = s$. Set $\Theta = \{d_j : \omega_j \in \Omega\}$. From condition (c) above we have:

$$|\Theta| := M \geq \frac{s-1}{8} \log\left(1 + \frac{p-1}{s-1}\right).$$

Further note that for any $d_i \neq d_j \in \Theta$:

$$\begin{aligned}
\|d_i - d_j\|_2^2 &= \frac{\epsilon^2}{(s-1)(1+\epsilon^2)} \|\omega_i - \omega_j\|^2 \\
&= \frac{\epsilon^2}{(s-1)(1+\epsilon^2)} d_H(\omega_i, \omega_j) \\
&= \frac{\epsilon^2}{2(1+\epsilon^2)} \geq \frac{\epsilon^2}{4}\,.
\end{aligned}$$

which proves that Θ is a $\epsilon/2$ packing set of S^{p-1}. On the other hand, from condition (c) above it is immediate that, for any $\omega_i \neq \omega_j \in \Omega$, we have $d_H(\omega_i, \omega_j) \leq 2s$. This implies that for any $d_i, d_j \in \Theta$:

$$\|d_i - d_j\|_2^2 = \frac{\epsilon^2}{(s-1)(1+\epsilon^2)} d_H(\omega_i, \omega_j) \leq 2\epsilon^2\,.$$

Now for each $d_i \in \Theta$ define the distribution P_{d_i} of (X, Y) as: $X \sim \mathcal{N}(0, I_p), \xi \sim \mathcal{N}(0,1)$, X is independent of ξ and:

$$Y \stackrel{d}{=} \mathbb{1}_{X^\top d_i > 0} + \xi\,.$$

Hence for any $d_i \neq d_j \in \Theta$, the Kullback-Liebler divergence between P_{d_i} and P_{d_j} is:

$$\begin{aligned} KL(P_{d_i}||P_{d_j}) &= \frac{1}{2}\mathbb{E}_X\left[\left(\mathbb{1}_{X^\top d_i>0} - \mathbb{1}_{X^\top d_j>0}\right)^2\right] \\ &= \mathbb{P}\left(\text{sign}(X^\top d_i) \neq \text{sign}(X^\top d_j)\right) \\ &\leq C\|d_i - d_j\|_2 \leq \epsilon\,\sqrt{2C^2}\,. \end{aligned}$$

for some universal constant C. Hence applying Fano's inequality we obtain:

$$\begin{aligned} \inf_{\hat{d}} \sup_{\mathcal{P}_d} \mathbb{E}\left[\left\|\hat{d} - d\right\|^2\right] &\geq \frac{\epsilon^2}{16}\left(1 - \frac{\frac{n}{M^2}\sum_{i,j:d_i,d_j\in\Theta} KL(P_{d_i}||P_{d_j}) + \log 2}{\log\left(|\Theta| - 1\right)}\right) \\ &\geq \frac{\epsilon^2}{16}\left(1 - \frac{n\epsilon\,\sqrt{2C^2} + \log 2}{\log\left(\frac{s-1}{8}\log\left(1 + \frac{p-1}{s-1}\right) - 1\right)}\right)\,. \end{aligned}$$

Taking $\epsilon = (s\log(1 + p/s))/n$ we conclude the proof.

Appendix B: Proof of supplementary lemmas

B.1. Proof of Lemma A.8

As ξ has symmetric distribution around origin, without loss of generality we can assume $\mu > 0$. Hence we have to establish the result for $0 < \mu k$. Note that difference $H_k(\xi + \mu) - H_k(\xi)$ can be decomposed into five terms, depending where ξ lies:

$$H_k(\xi+\mu) - H_k(\xi) = \begin{cases} \frac{1}{2}\left[(\xi+\mu)^2 - \xi^2\right]\,, & \text{if } -k \leq \xi \leq k - \mu \\ \frac{1}{2}(\xi+\mu)^2 - k\left(|\xi| - \frac{k}{2}\right)\,, & \text{if } -k-\mu \leq \xi \leq -k \\ k\left(|\xi+\mu| - \frac{k}{2}\right) - \frac{\xi^2}{2}\,, & \text{if } k-\mu \leq \xi \leq k \\ K\left(|\xi+\mu| - |\xi|\right)\,, & \text{if } \xi > k \text{ or } \xi < -k-\mu \end{cases} \tag{B.1}$$

Now we inspect the regions individually. Note that when $-k - \mu \leq \xi \leq -k$, we have:

$$\begin{aligned} H_k(\xi+\mu) - H_k(\xi) &= \frac{1}{2}(\xi+\mu)^2 - k\left(|\xi| - \frac{k}{2}\right) \\ &= \frac{1}{2}(\xi+\mu)^2 + k\left(\xi + \frac{k}{2}\right) \qquad [\because |\xi| = -\xi] \end{aligned}$$

$$= \frac{\xi^2}{2} + (\mu + k)\xi + \frac{\mu^2 + k^2}{2}$$
$$= \frac{1}{2}(\xi + \mu + k)^2 - \mu k$$

When $k - \mu \leq \xi \leq k$:

$$\begin{aligned} H_k(\xi + \mu) - H_k(\xi) &= k\left(|\xi + \mu| - \frac{k}{2}\right) - \frac{\xi^2}{2} \\ &= k\left((\xi + \mu) - \frac{k}{2}\right) - \frac{\xi^2}{2} \\ &= -\frac{\xi^2}{2} + k\xi - \frac{k^2}{2} + k\mu \\ &= -\frac{1}{2}(\xi - k)^2 + k\mu \end{aligned}$$

Also, we have:

$$k\left(|\xi + \mu| - |\xi|\right) = \begin{cases} k\mu\,, & \text{if } \xi > k \\ -k\mu\,, & \text{if } \xi < -k - \mu\,. \end{cases}$$

Hence we can modify equation (B.1) as:

$$H_k(\xi + \mu) - H_k(\xi) = \begin{cases} \mu\xi + \frac{\mu^2}{2}\,, & \text{if } -k \leq \xi \leq k - \mu \\ \frac{1}{2}(\xi + \mu + k)^2 - \mu k\,, & \text{if } -k - \mu \leq \xi \leq -k \\ -\frac{1}{2}(\xi - k)^2 + k\mu\,, & \text{if } k - \mu \leq \xi \leq k \\ k\mu\,, & \text{if } \xi > k \\ -k\mu\,, & \text{if } \xi < -k - \mu\,. \end{cases} \tag{B.2}$$

Note that the term $-\mu k$ is active on the region $\xi \leq -k$ and μk is active on the region $\xi \geq k - \mu$. From the symmetry of the distribution of ξ, this effect of $-\mu k$ and μk on the region $(-\infty, -k)$ and (k, ∞) will cancel each other upon taking expectation and the effect of μk on $(k - \mu, k)$ will remain. Hence we have:

$$\begin{aligned} &\mathbb{E}\left[H_k(\xi + \mu) - H_k(\xi)\right] \\ &= \frac{\mu^2}{2}\mathbb{P}\left(-k \leq \xi \leq k - \mu\right) + \mu\mathbb{E}\left[\xi \mathbb{1}_{-k \leq \xi \leq k - \mu}\right] \\ &\quad + \mathbb{E}\left[\frac{1}{2}(\xi + \mu + k)^2 \mathbb{1}_{-k - \mu \leq \xi \leq -k}\right] + \mathbb{E}\left[-\frac{1}{2}(\xi - k)^2 \mathbb{1}_{k - \mu \leq \xi \leq k}\right] \\ &= \frac{\mu^2}{2}\mathbb{P}\left(-k \leq \xi \leq k - \mu\right) - \mu\mathbb{E}\left[\xi \mathbb{1}_{k - \mu \leq \xi \leq k}\right] \\ &\quad + \mathbb{E}\left[\frac{1}{2}(\xi + \mu + k)^2 \mathbb{1}_{-k - \mu \leq \xi \leq -k}\right] + \mathbb{E}\left[\left(\mu k - \frac{1}{2}(\xi - k)^2\right) \mathbb{1}_{k - \mu \leq \xi \leq k}\right] \\ &= \frac{\mu^2}{2}\mathbb{P}\left(-k \leq \xi \leq k - \mu\right) + \mathbb{E}\left[\left(\mu k - \mu\xi - \frac{1}{2}(\xi - k)^2\right) \mathbb{1}_{k - \mu \leq \xi \leq k}\right] \end{aligned}$$

$$
\begin{aligned}
&\quad + \mathbb{E}\left[\frac{1}{2}\left(\xi+\mu+k\right)^2 \mathbb{1}_{-k-\mu\le\xi\le-k}\right] \\
&\ge \frac{\mu^2}{2}\mathbb{P}\left(-k\le\xi\le k-\mu\right) + \mathbb{E}\left[\left(\mu k-\mu\xi-\frac{1}{2}\left(\xi-k\right)^2\right)\mathbb{1}_{k-\mu\le\xi\le k}\right] \\
&\ge \frac{\mu^2}{2}\mathbb{P}\left(-k\le\xi\le k-\mu\right)
\end{aligned}
$$

where the last inequality follows from the fact:

$$
f(\xi) = \mu k - \mu\xi - \frac{1}{2}\left(\xi-k\right)^2 \ge 0 \quad \forall \quad \xi\in[k-\mu,k]\ .
$$

observing the fact that:

$$
\begin{aligned}
\mathbb{E}\left[\tilde{H}_k(\xi+\mu)-\tilde{H}_k(\xi)\right] &= \frac{k+1}{k}\mathbb{E}\left[H_k(\xi+\mu)-H_k(\xi)\right] \\
&\ge \mathbb{E}\left[H_k(\xi+\mu)-H_k(\xi)\right]
\end{aligned}
$$

we complete the proof for all $k>0$. Now for $k=0$ for $0<\mu<\delta$,

$$
\begin{aligned}
&\mathbb{E}\left[|\xi+\mu|-|\xi|\right] \\
&= -\mu\mathbb{P}(\xi\le-\mu)+\mu\mathbb{P}(\xi>0)+\mathbb{E}\left[(2\xi+\mu)\mathbb{1}_{-\mu\le\xi\le0}\right] \\
&= \mu\mathbb{P}(0\le\xi\le\mu)+\mathbb{E}\left[(-2\xi+\mu)\mathbb{1}_{0\le\xi\le\mu}\right] \\
&= \mathbb{E}\left[(-2\xi+2\mu)\mathbb{1}_{0\le\xi\le\mu}\right] \\
&= 2\int_0^\mu(\mu-x)f_\xi(x)\ dx \\
&\ge f_\xi(0)\int_0^\mu(\mu-x)\ dx \\
&= \frac{\mu^2}{2}f_\xi(0)\,.
\end{aligned}
$$

This completes the proof.

B.2. Proof of Lemma A.1

Proof. Although we assume continuous steps, our proof can be certainly extended to the case when S_n takes value 0 with positive probability. The proof critically uses Theorem 4 of Chapter 12 of Volume 2 of [16]. To keep the notational similarity with the book, define:

$$
q_n = \mathbb{P}\left(\max_{1\le i\le n} S_i < 0\right)
$$

and the corresponding generating function $q(s)$ as:

$$q(s) = 1 + \sum_{n=1}^{\infty} s^n q_n \,.$$

Then from equation (7.22) of Theorem 4, Chapter 12, Vol. 2 of [16] we have:

$$\log q(s) = \sum_{n=1}^{\infty} \frac{s^n}{n} \mathbb{P}\left(S_n < 0\right) := f(s) \iff q(s) = e^{f(s)} \,. \tag{B.3}$$

Now we need a lower bound on q_n. Note that from the property of the generating function we have:

$$q_n = n! q^{(n)}(0) \,.$$

On the other hand from Faa di Bruno's formula:

$$\begin{aligned}
& q^{(n)}(0) \\
&= \left. \frac{d^n}{ds^n} e^{f(s)} \right|_{s=0} \\
&= \left. \left[e^{f(s)} \sum \frac{n!}{m_1! 1!^{m_2} m_1! 2!^{m_2} \cdots m_n! n!^{m_n}} \Pi_{j=1}^{n} \left(f^{(j)}(s) \right)^{m_j} \right] \right|_{s=0} \\
&= \left. \left[e^{f(s)} \sum \frac{n!}{m_1! m_2! \cdots m_n!} \Pi_{j=1}^{n} \left(\frac{f^{(j)}(s)}{j!} \right)^{m_j} \right] \right|_{s=0} \\
&= \sum \frac{n!}{m_1! m_2! \cdots m_n!} \Pi_{j=1}^{n} \left(\frac{f^{(j)}(0)}{j!} \right)^{m_j}
\end{aligned} \tag{B.4}$$

where the sum runs over all the sequences $\{m_j\}_{j=1}^{n}$ such that:

$$\sum_{i=1}^{n} i m_i = n \,.$$

Note that $f^{(j)}(0)$ is non-negative for j. Hence using only one sequence with $m_1 = \cdots = m_{n-1} = 0$ and $m_n = 1$, equation (B.4) can lower bounded as:

$$\left. \frac{d^n}{ds^n} e^{f(s)} \right|_{s=0} \geq f^{(n)}(0) \,.$$

On the from the expression of $f(s)$ from equation (B.3) it is immediate that:

$$f^{(n)}(0) = n! \frac{1}{n} \mathbb{P}\left(S_n < 0\right) \,.$$

Combining our findings we have:

$$\begin{aligned}
n! q_n = q^{(n)}(0) &= \left. \frac{d^n}{ds^n} e^{f(s)} \right|_{s=0} \\
&\geq f^{(n)}(0) \\
&= n! \frac{1}{n} \mathbb{P}\left(S_n < 0\right) \,.
\end{aligned}$$

This immediately implies $q_n \geq \mathbb{P}\left(S_n > 0\right)/n$ which completes our proof. □

B.3. Proof of Lemma A.2

From the definition of distribution of ξ we have:

$$F_\xi(t) = \frac{\frac{1}{2} + t^\gamma}{1 + t^\gamma} \qquad t \geq 0\,.$$

and

$$F_\xi(-t) = 1 - F_\xi(t) = \frac{1}{2(1+t^\gamma)}\,.$$

Hence it is immediate that for any $t_0 > 0$:

$$\frac{1}{2(1+t_0^{-\gamma})} \leq \sup_{t \geq t_0} t^\gamma \bar{F}_\xi(t) = \sup_{t \geq t_0} \frac{t^\gamma}{2(1+t^\gamma)} \leq \frac{1}{2} \tag{B.5}$$

For any fixed $k \geq 1$:

$$\begin{aligned}
\mathbb{P}\left(M \geq k\right) &= \sum_{j \geq k} \mathbb{P}\left(M = j\right) \\
&= \sum_{j \geq k} \mathbb{P}\left(S_i > S_j \;\forall\; 0 \leq i \leq j-1, S_i > S_j \;\forall\; i \geq j+1\right) \\
&= \mathbb{P}\left(S_1 > 0, S_2 > 0, \dots\right) \sum_{j \geq k} \mathbb{P}\left(\max_{1 \leq i \leq j} S_i < 0\right) \\
&= p^* \sum_{j \geq k} \mathbb{P}\left(\max_{1 \leq i \leq j} S_j < 0\right) \qquad [p^* = P\left(S_1 > 0, S_2 > 0, \dots\right)] \\
&\geq p^* \sum_{j \geq k} \frac{1}{j} P\left(S_j \leq 0\right)
\end{aligned} \tag{B.6}$$

where the last inequality uses Lemma A.1. From the symmetry of the distribution of ξ we have:

$$\mathbb{P}\left(S_j \leq 0\right) = \mathbb{P}\left(\sum_{i=1}^{j} \xi_j \leq -j\mu\right) = \mathbb{P}\left(\sum_{i=1}^{j} \xi_j > j\mu\right)\,.$$

Set $a_j = j^{1/\gamma}$. Define the event A_i as:

$$A_i = \{\xi_i > j\mu + (j-1)a_j,\; \xi_l \in [-a_j, j\mu) \;\forall\; 1 \leq l \neq i \leq j\}$$

Clearly $\{A_i\}'s$ are disjoint events and

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{j} \xi_j > j\mu\right) &\geq \mathbb{P}\left(\cup_{i=1}^{j} A_i\right) \\
&= \sum_{i=1}^{j} \mathbb{P}(A_i)
\end{aligned}$$

$$= j\bar{F}\left(j\mu + (j-1)a_j\right)\left(F[-a_j, j\mu)\right)^{j-1}$$
$$= j\bar{F}\left(j\mu + (j-1)a_j\right)\left(1 - \bar{F}(a_j) - \bar{F}(j\mu)\right)^{j-1} \tag{B.7}$$

Next note that, $j\mu + (j-1)a_j \geq \mu$ for all $j \geq 1$. Therefore from equation (B.5) we have for all $j \geq 1$:

$$\bar{F}\left(j\mu + (j-1)a_j\right) \geq \frac{\left(j\mu + (j-1)a_j\right)^{-\gamma}}{2(1+\mu^{-\gamma})},$$

which further implies:

$$\begin{aligned}
j \times \bar{F}\left(j\mu + (j-1)a_j\right) &\geq \frac{1}{2(1+\mu^{-\alpha})} \frac{j}{\left(j\mu + (j-1)a_j\right)^{\gamma}} \\
&= \frac{1}{2(1+\mu^{-\gamma})} \frac{j}{\left(j\mu + (j-1)j^{1/\alpha}\right)^{\gamma}} \\
&= \frac{1}{2(1+\mu^{-\gamma})} \frac{j}{j^{\gamma+1}\left(j^{-1/\gamma}\mu + \left(1 - \frac{1}{j}\right)\right)^{\gamma}} \\
&= \frac{1}{j^{\gamma}} \frac{1}{2(1+\mu^{-\gamma})} \frac{1}{\left(j^{-1/\gamma}\mu + \left(1 - \frac{1}{j}\right)\right)^{\gamma}} \\
&\geq \frac{1}{j^{\gamma}} \frac{1}{2(1+\mu^{-\gamma})} \frac{1}{(\mu+1)^{\gamma}} := \frac{c_1}{j^{\gamma}}.
\end{aligned}$$

Next observe that $(j\mu)^{\gamma} \geq j$ for all $j \geq 1$ if $\mu > 1$ or for all $j \geq \mu^{-\gamma/(\gamma-1)}$ if $\mu \leq 1$. Using this in equation (B.24) we have for all $j \geq 1 \vee \lceil \mu^{-\gamma/(\gamma-1)} \rceil$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{j} \xi_j > j\mu\right) &\geq \frac{c_1}{j^{\gamma}}\left[\left(1 - \bar{F}(a_j) - \bar{F}(j\mu)\right)^{j-1}\right] \\
&\geq \frac{c_1}{j^{\gamma}} \times \left(1 - \frac{1}{2(1+a_j^{\alpha})} - \frac{1}{2(1+(j\mu)^{\alpha})}\right)^{j-1} \\
&= \frac{c_1}{j^{\gamma}} \times \left(1 - \frac{1}{2(1+j)} - \frac{1}{2(1+(j\mu)^{\alpha})}\right)^{j-1} \\
&\geq \frac{c_1}{j^{\gamma}} \times \left(1 - \frac{1}{(1+j)}\right)^{j-1} \\
&\geq \frac{c_1}{j^{\gamma}} \times \inf_{x \geq 1}\left(1 - \frac{1}{(1+x)}\right)^{x-1} \\
&:= \frac{c_1 c_2}{j^{\gamma}}
\end{aligned} \tag{B.8}$$

Using this in equation (B.23) we obtain:

$$P\left(M \geq k\right) \geq p^* \sum_{j \geq k} \frac{1}{j} P\left(S_j \leq 0\right)$$

$$
\begin{aligned}
&= p^* \sum_{j \geq k} \frac{1}{j} \left(\sum_{i=1}^{j} \xi_j > j\mu \right) \\
&\geq c_1 c_2 \times p^* \times \sum_{j \geq k} j^{-(\gamma+1)} \\
&\geq c_1 c_2 \times p^* \times \int_k^{\infty} x^{-(\gamma+1)} \, dx \\
&\geq c_1 c_2 \times p^* \times \frac{1}{\gamma k^{\gamma}} .
\end{aligned}
$$

This completes the proof of lower bound.

B.4. Proof of Lemma A.3

We have, by symmetry:

$$\mathbb{P}\left(|M_{ts}| > k\right) = \mathbb{P}\left(M_{ts} > k\right) + \mathbb{P}\left(M_{ts} < -k\right) = 2\mathbb{P}\left(M_{ts} > k\right) .$$

Hence, by virtue of Lemma A.2, all we need to show is:

$$P(M_{ts} = k) \geq p^* \mathbb{P}(M_{os} = k) .$$

Towards that end:

$$
\begin{aligned}
\mathbb{P}\left(M_{ts} = k\right) &= \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1, S_k \leq S_i \ \forall \ i \geq k+1, S_k \leq \inf_{j \geq 1} S_{-j} \right) \\
&= \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1, S_k \leq \inf_{j \geq 1} S_{-j} \right) \mathbb{P}\left(S_i \geq 0 \ \forall \ i \geq 1 \right) \\
&= p^* \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1, S_k \leq \inf_{j \geq 1} S_{-j} \right) \\
&\geq p^* \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1, S_k \leq \inf_{j \geq 1} S_{-j} \mid \inf_{j \geq 1} S_{-j} > 0 \right) \\
&\qquad \times \mathbb{P}\left(\inf_{j \geq 1} S_{-j} > 0 \right) \\
&= p^* \mathbb{P}\left(\inf_{j \geq 1} S_{-j} > 0 \right) \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1 \right) \\
&= \mathbb{P}\left(\inf_{j \geq 1} S_{-j} > 0 \right) \mathbb{P}\left(M_{os} = k \right) \\
&= p^* \mathbb{P}\left(M_{os} = k \right) .
\end{aligned}
$$

where the last equality follows from the fact:

$$\mathbb{P}\left(M_{os} = k\right) = p^* \mathbb{P}\left(S_K \leq S_i \ \forall \ 0 \leq i \leq k-1\right)$$

This completes the proof.

B.5. Proof of Lemma A.4

Using same line of arguments as in Corollary 1:

$$\begin{aligned}
&\mathbb{P}\left(M_{ts,CPP}=x\right)\\
&=\sum_{k=0}^{\infty}\mathbb{P}\left(M_{ts,CPP}=x \mid N_1(x)=k\right)\mathbb{P}(N_1(x)=k)\\
&=\sum_{k=0}^{\infty}\mathbb{P}\left(M_{ts}=k\right)\mathbb{P}(N_1(x)=k)\\
&\geq p^*\sum_{k=0}^{\infty}\mathbb{P}\left(M_{os}=k\right)\mathbb{P}(N_1(x)=k)\\
&=\mathbb{P}\left(M_{os,CPP}=x\right)
\end{aligned}$$

Hence to establish Corollary A.4, all we need show:

$$\mathbb{P}\left(M_{os,CPP}>x\right)\geq\frac{c_0}{2f_X^{\gamma}(d_0)}x^{-\gamma}$$

for all large x, where M is the argmin of one sided compound Poisson process, namely the minimizer of the following:

$$X_+(t)=\sum_{i=1}^{N_1(t)}X_i\,,\quad t\in\mathbb{R}^+\,.$$

Now we have:

$$\begin{aligned}
P(M_{os,CPP}>x)&=\sum_{k=0}^{\infty}\mathbb{P}\left(M_{os,CPP}>x \mid N_1(x)=k\right)\mathbb{P}\left(N_1(x)=k\right)\\
&=\sum_{k=0}^{\infty}\mathbb{P}\left(\operatorname*{argmin}_{i\geq 0}S_i>k\right)\mathbb{P}\left(N_1(x)=k\right)\\
&\geq\sum_{k=k_0}^{\infty}\mathbb{P}\left(\operatorname*{argmin}_{i\geq 0}S_i>k\right)\mathbb{P}\left(N_1(x)=k\right)\\
&\geq c_0\sum_{k=k_0}^{\infty}k^{-\gamma}\mathbb{P}\left(N_1(x)=k\right)\\
&=c_0\sum_{k=k_0}^{\infty}\frac{e^{-\Lambda(x)}\Lambda(x)^k}{k!k^{\gamma}}\\
&\geq c_0\sum_{k=k_0}^{\infty}\frac{e^{-\Lambda(x)}\Lambda(x)^k}{k!(k+1)(k+2)\ldots(k+\gamma)}\\
&=c_0\sum_{k=k_0}^{\infty}\frac{e^{-\Lambda(x)}\Lambda(x)^k}{(k+\gamma)!}
\end{aligned}$$

$$
\begin{aligned}
&= c_0 \Lambda(x)^{-\gamma} \sum_{k=k_0+\alpha}^{\infty} \frac{e^{-\Lambda(x)} \Lambda(x)^k}{k!} \\
&= c_0 \Lambda(x)^{-\gamma} \mathbb{P}\left(N_1(x) \geq k_0 + \alpha\right) \\
&\geq \frac{c_0}{2} \Lambda(x)^{-\gamma} = \frac{c_0}{2 f_X^{\gamma}(d_0)} x^{-\gamma}
\end{aligned}
$$

where the last inequality is valid as long as $\mathsf{med}\left(N_1(X)\right) \geq k_0 + \gamma$. From [2], we know as $N_1(x) \sim Poisson(x f_X(d_0))$, we have $\mathsf{med}\left(N_1(x)\right) \geq x f_X(d_0) - \log 2$. Hence the inequality is valid as long as $x \geq (k_0 + \gamma + \log 2)/f_X(d_0)$. From This completes the proof.

B.6. Proof of Lemma A.5

Proof. As per our model description, all the parallel change point processes are i.i.d. Therefore $n(\hat{d}_i - d_{0,i})$ has same distribution across $1 \leq i \leq m$. Therefore, we henceforth define F_n to be the distribution of $n(\hat{d} - d_0)$ and drop i from subscript. From the definition of change point estimator, we have:

$$
\begin{aligned}
n(\hat{d} - d_0) &= \text{mid argmin}_t \sum_{i=1}^{n} \left(\xi_i + \frac{1}{2}\right) \left\{\mathbb{1}_{d_0 < X_i \leq d_0 + \frac{t}{n}}\right\} \\
&\qquad + \text{mid argmin}_t \sum_{i=1}^{n} \left(-\xi_i + \frac{1}{2}\right) \left\{\mathbb{1}_{d_0 + \frac{t}{n} < X_i \leq d_0}\right\} \\
&= \text{mid argmin}_t \sum_{i=1}^{N_{n,+}(t)} \left(\xi_i + \frac{1}{2}\right) \mathbb{1}_{t \geq 0} + \sum_{i=1}^{N_{n,-}(t)} \left(-\xi_i + \frac{1}{2}\right) \mathbb{1}_{t < 0} \,.
\end{aligned}
$$

Here the count processes $N_{n,+}(t)$ and $N_{n,-}(t)$ are defined as follows: For $t \geq 0$,

$$
N_{n,+}(t) = \sum_{i=0}^{n} \mathbb{1}_{d_0 \leq X_i \leq d_0 + \frac{t}{n}} \sim \text{Bin}\left(n, F_X\left(d_0 + \frac{t}{n}\right) - F_X(t)\right)
$$

and for $t < 0$,

$$
N_{n,-}(t) = \sum_{i=0}^{n} \mathbb{1}_{d_0 + \frac{t}{n} \leq X_i \leq d_0} \sim \text{Bin}\left(n, F_X(t) - F_X\left(d_0 + \frac{t}{n}\right)\right) .
$$

These processes can be though as finite sample approximation of Compound Poisson Process, where we have approximated the Poisson random variables by Binomial random variables. It is immediate that:

$$
\begin{aligned}
N_{n,+}(t) &\overset{\mathscr{L}}{\Longrightarrow} \text{Pois}\left(t f(d_0)\right) , \\
N_{n,-}(t) &\overset{\mathscr{L}}{\Longrightarrow} \text{Pois}\left(-t f(d_0)\right) .
\end{aligned}
$$

We name the process as *compound Binomial process* and henceforth denote by CBP:

$$CBP(t) = \sum_{i=1}^{N_{n,+}(t)} \left(\xi_i + \frac{1}{2}\right) \mathbb{1}_{t\geq 0} + \sum_{i=1}^{N_{n,-}(t)} \left(-\xi_i + \frac{1}{2}\right) \mathbb{1}_{t<0} \tag{B.9}$$

Hence we work with the smallest argmin instead of mid-argmin just for some technical simplicity, but all of the following analysis is valid for mid-argmin also. As will be evident later, the thickness of the tail of the distribution F_n (the distribution of $n(\hat{d} - d_0)$) is closely related to the tail of the minimizer of a random walk with finitely many steps. Therefore, we start by establishing a lower bound on the tail of a n-step random walk $\{S_i\}_{i=0}^n$ with the usual convention $S_0 = 0$ and step distribution $X_i \stackrel{d}{=} \xi_i + 1/2$. Let Z_n be the minimizer of this random walk. The random variable Z_n is supported on $\{0, 1, \ldots, n\}$. Then for any $0 \leq k \leq n-1$:

$$\begin{aligned}
\mathbb{P}(Z_n > k) &= \sum_{j=k+1}^{n} \mathbb{P}(Z_n = j) \\
&= \sum_{j=k+1}^{n} \mathbb{P}(S_i > S_j \ \forall\ 0 \leq i \leq j-1, S_i > S_j \ \forall\ j+1 \leq i \leq n) \\
&= \sum_{j=k+1}^{n} P(S_i < 0 \ \forall\ 1 \leq i \leq j)\, \mathbb{P}(S_i > 0 \ \forall\ 1 \leq i \leq n-j) \\
&\geq \mathbb{P}(S_i > 0 \ \forall\ 1 \leq i < \infty) \sum_{j=k+1}^{n} \mathbb{P}\left(\max_{1\leq i\leq j} S_i < 0\right) \\
&= p^* \sum_{j=k+1}^{n} \mathbb{P}\left(\max_{1\leq i\leq j} S_j < 0\right) \qquad \left[p^* = \mathbb{P}\left(\min_{1\leq i<\infty} S_i > 0\right)\right] \\
&\geq p^* \sum_{j=k+1}^{n} \frac{1}{j} \mathbb{P}(S_j \leq 0)
\end{aligned} \tag{B.10}$$

From equation (B.25) in the proof of Lemma A.2 we conclude for $j \geq 2^{\gamma/(\gamma-1)} := k_0$:

$$\mathbb{P}(S_j \leq 0) \geq \frac{c_1 c_2}{j^\gamma}\,.$$

Using the above bound in equation (B.10) we conclude:

$$\begin{aligned}
\mathbb{P}(Z_n > k) &\geq p^* \sum_{j=k+1}^{n} \frac{1}{j} \mathbb{P}(S_j \leq 0) \\
&\geq p^* \sum_{j=k+1}^{n} \frac{1}{j} \frac{c_1 c_2}{j^\gamma}
\end{aligned}$$

$$
\begin{aligned}
&= c_1 c_2 p^* \sum_{j=k+1}^{n} \frac{1}{j^{\gamma+1}} \\
&\geq c_1 c_2 p^* \int_{k+1}^{n+1} x^{-(\gamma+1)}\, dx \quad \text{[Riemann integral lower bound]} \\
&= \frac{c_1 c_2 p^*}{\gamma} \left[\frac{1}{(k+1)^{\gamma}} - \frac{1}{(n+1)^{\gamma}} \right] \qquad \text{(B.11)}
\end{aligned}
$$

Now we go back to the random variable of interest $n(\hat{d}-d_0)$. Let $X_{(i)}$ denotes the i^{th} order statistics of $\{X_i\}_{i\leq n}$. If $X_{(i)} < d_0 < X_{(i+1)}$, then from the definition of $n(\hat{d}-d_0)$, we have a random walk with i steps on the negative axis and a random walk with $n-i$ steps on the positive axis. Therefore the number of steps of random walk on either side of origin is equal to the number of $X_i's$ on the corresponding side of d_0. Denote by R_n (and respectively L_n), the number of $X_i's$ greater than d_0 (respectively less than d_0). Hence $R_n \sim \text{Bin}(n, \bar{F}_X(d_0))$ and $L_n \sim \text{Bin}(n, F_X(d_0))$ with $R_n + L_n = n$. Then we have for any $x > 0$:

$$
\begin{aligned}
&\mathbb{P}\left(n(\hat{d}-d_0) > x\right) \\
&= \sum_{r=0}^{n} \sum_{k=0}^{r} \Big\{ \mathbb{P}\left(n(\hat{d}-d_0) > x \mid R_n = r, N_{n,+}(x) = k\right) \\
&\qquad\qquad \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \Big\}
\end{aligned}
$$

Given $R_n = r$, we have a two sided random walk, with r steps on the positive real line $n-r$ steps on the negative real line. Therefore, the event $n(\hat{d}-d_0) > x$ given $N_{n,+}(x) = k$ and $R_n = r$ is equivalent to the event that in a two sided random walks with r steps on the right and $n-r$ steps on the left, the argmin is on the right and it happens after k steps. More precisely, if we denote by $S_0 \equiv 0, S_1, \ldots, S_r$ to be the random walk on the right side with step distribution $(\xi + 1/2)$ and $S'_0 \equiv 0, S'_1, \ldots, S'_{n-r}$ to be random walk on the left with step distribution $(-\xi + 1/2)$, then the above event corresponds that this two sided random walk is minimized at S_j for some $k+1 \leq j \leq r$. Therefore we write:

$$
\begin{aligned}
&\mathbb{P}\left(n(\hat{d}-d_0) > x\right) \\
&= \sum_{r=0}^{n} \sum_{k=0}^{r} \Big[\mathbb{P}\left(n(\hat{d}-d_0) > x \mid R_n = r, N_{n,+}(x) = k\right) \qquad \text{(B.12)} \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \Big] \\
&= \sum_{r=0}^{n} \sum_{k=0}^{r} [\mathbb{P}\left(\text{argmin of twosided RW } > k\right) \qquad \text{(B.13)} \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \qquad \text{(B.14)}
\end{aligned}
$$

Next, we obtain a lower bound on the tail of the minimizer of the two sided random walk. Note that, we have already established a lower bound on the

tail of the minimizer of a one sided random walk in equation (B.11), which we exploit here to get a lower bound on the tail of the minimizer of this two-sided incarnation:

$$
\begin{aligned}
&\mathbb{P}\left(\text{argmin twosided RW } > k\right) \\
&= \sum_{j=k+1}^{r} \mathbb{P}\left(\text{argmin twosided RW } = j\right) \\
&= \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < S_0, \ldots, S_j < S_{j-1}, S_j < S_{j+1}, \ldots, S_j < S_r, S_j < \min_{1\le i\le n-r} S'_i\right) \\
&= \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < S_0, \ldots, S_j < S_{j-1}, S_j < \min_{1\le i\le n-r} S'_i\right) \mathbb{P}(S_1 > 0, \ldots, S_{r-j} > 0) \\
&\ge \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < S_0, \ldots, S_j < S_{j-1}, S_j < \min_{1\le i\le n-r} S'_i \mid \min_{1\le i\le n-r} S'_i > 0\right) \times \\
&\qquad\qquad \mathbb{P}\left(\min_{1\le i\le n-r} S'_i > 0\right) \mathbb{P}(S_1 > 0, \ldots, S_{r-j} > 0) \\
&\ge p^* \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < S_0, \ldots, S_j < S_{j-1}, S_j < \min_{1\le i\le n-r} S'_i \mid \min_{1\le i\le n-r} S'_i > 0\right) \times \\
&\qquad\qquad \mathbb{P}(S_1 > 0, \ldots, S_{r-j} > 0) \\
&= p^* \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < S_0, \ldots, S_j < S_{j-1}\right) \mathbb{P}(S_1 > 0, \ldots, S_{r-j} > 0) \\
&= p^* \mathbb{P}\left(\text{argmin one-sided RW with length } r > k\right) \\
&\ge \frac{c_1 c_2 (p^*)^2}{\gamma} \left[\frac{1}{(k+1)^\gamma} - \frac{1}{(r+1)^\gamma}\right] \qquad \text{[From equation (B.11)]}.
\end{aligned}
$$

For the rest of the calculation we assume γ (the number of finite moments of the error distribution ξ) is an integer, as all our calculation is valid by replacing γ by $\lfloor\gamma\rfloor$. Define a success probability $p_{x,n}$ as:

$$
p_{x,n} = \mathbb{P}\left(X \in \left(d_0 + \frac{x}{n}, d_0\right) \mid X > d_0\right) = \frac{F_X\left(d_0 + \frac{x}{n}, d_0\right) - F_X(d_0)}{1 - F_X(d_0)}.
$$

Therefore it is immediate that:

$$
N_{n,+}(x) \mid R_n = r \sim \text{Bin}\left(r, p_{x,n}\right).
$$

As per our assumption F_X has continuous density f_X with $f_X(d_0) > 0$. Therefore, there exists $\delta_1 > 0$ such that $f_X(t) > f_X(d_0)/2$ for $|t - d_0| \le \delta_1$. Hence,

for any $0 \leq x \leq n\delta_1$, we have:

$$p_{x,n} \geq \frac{x}{n} \times \frac{f_X(d_0)}{2(1 - F_X(d_0))}\,. \tag{B.15}$$

On the other, let $f_{\max}$ be the upper bound on f_X on the entire $\mathbb{R}$. Then, again from the mean value theorem, we have:

$$p_{x,n} \leq \frac{x}{n} \times \frac{f_{\max}}{2(1 - F_X(d_0))}\,. \tag{B.16}$$

Therefore combining equations (B.15) and (B.16), we have:

$$\frac{x}{n} \times \frac{f_X(d_0)}{2(1 - F_X(d_0))} \leq p_{x,n} \leq \frac{x}{n} \times \frac{f_{\max}}{2(1 - F_X(d_0))}\,. \tag{B.17}$$

We will use the above relations in our rest of the calculation. Going back to equation (B.13) we have:

$$\begin{aligned}
&\mathbb{P}\left(n(\hat{d} - d_0) > x\right) \\
&= \sum_{r=1}^{n}\sum_{k=0}^{r} [\mathbb{P}\left(\text{argmin twosided RW } > k\right) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \\
&= \sum_{r=k_0}^{n}\sum_{k=k_0}^{r} [\mathbb{P}\left(\text{argmin twosided RW } > k\right) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \\
&\geq \frac{c_1 c_2 (p^*)^2}{\gamma} \sum_{r=1}^{n}\sum_{k=0}^{r} \left\{ \left[\frac{1}{(k+1)^\gamma} - \frac{1}{(r+1)^\gamma} \right] \right. \\
&\qquad\qquad \left. \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \right\} \\
&= \frac{c_1 c_2 (p^*)^2}{\gamma} \left[\sum_{r=k_0}^{n}\sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \right. \\
&\qquad\qquad \left. - \sum_{r=k_0}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(k_0 \leq N_{n,+}(x) < r \mid R_n = r)\mathbb{P}(R_n = r) \right] \\
&= \frac{c_1 c_2 (p^*)^2}{\gamma} \left[\sum_{r=k_0}^{n}\sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \right. \\
&\qquad\qquad \left. - \sum_{r=1}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(R_n = r) \right]
\end{aligned}$$

$$= \frac{c_1 c_2 (p^*)^2}{\gamma} \left[\sum_{r=k_0}^{n} \sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \binom{r}{k} p_{x,n}^k (1-p_{x,n})^{r-k} \mathbb{P}(R_n = r) \right.$$
$$\left. - \sum_{r=1}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(R_n = r) \right]$$
$$\geq \frac{c_1 c_2 (p^*)^2}{\gamma} \left[\sum_{r=k_0}^{n} \left\{ \sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \binom{r}{k} p_{x,n}^k (1-p_{x,n})^{r-k} \right\} \mathbb{P}(R_n = r) \right.$$
$$\left. - \sum_{r=1}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(R_n = r) \right] \tag{B.18}$$

where the last inequality is obtained by replacing $1 - p_{x,n}$ by its upper bound 1. The inner sum of the above equation can be analyzed as follows:

$$\sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \binom{r}{k} p_{x,n}^k (1-p_{x,n})^{r-k}$$
$$= \sum_{k=k_0}^{r} \frac{1}{(k+1)^\gamma} \frac{r!}{k!(r-k)!} p_{x,n}^k (1-p_{x,n})^{r-k}$$
$$\geq \sum_{k=k_0}^{r} \frac{1}{(k+1)(k+2)\dots(k+\gamma)} \frac{r!}{k!(r-k)!} p_{x,n}^k (1-p_{x,n})^{r-k}$$
$$\geq \frac{p_{x,n}^{-\gamma}}{(r+1)(r+2)\dots(r+\gamma)} \sum_{k=k_0}^{r} \frac{(r+\gamma)!}{(k+\gamma)!(r-k)!} p_{x,n}^{k+\gamma} (1-p_{x,n})^{r-k}$$
$$= \frac{p_{x,n}^{-\gamma}}{(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}\left(\mathrm{Bin}(r+\gamma, p_{x,n}) \geq k_0 + \gamma\right) .$$

Putting this back into equation (B.18) we obtain:

$$\mathbb{P}\left(n(\hat{d} - d_0) > x \right)$$
$$\geq \frac{c_1 c_2 (p^*)^2}{\gamma} p_{x,n}^{-\gamma} \left[\sum_{r=k_0}^{n} \frac{\mathbb{P}\left(\mathrm{Bin}(r+\gamma, p_{x,n}) \geq k_0 + \gamma\right)}{(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}(R_n = r) \right.$$
$$\left. - p_{x,n}^{\gamma} \sum_{r=1}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(R_n = r) \right] \tag{B.19}$$

Now from the properties of the inverse moments of the binomial distribution (see e.g. [11]) we have:

$$\sum_{r=1}^{n} \frac{1}{(r+1)^\gamma} \mathbb{P}(R_n = r) \leq C \left(n \bar{F}_X(d_0) \right)^{-\gamma} \qquad [\text{Recall } R_n \sim \mathrm{Bin}(n, \bar{F}_X(d_0))] .$$

On the other hand we have for all $r \geq n\bar{F}_X(d_0)$:

$$\begin{aligned}(r+\gamma)p_{x,n} &\geq (n\bar{F}_X(d_0)+\gamma) \times \frac{x}{n} \times \frac{f_X(d_0)}{2(1-F_X(d_0))} \qquad \text{[Equation (B.15)]} \\ &\geq \left(\bar{F}_X(d_0)+\frac{\gamma}{n}\right) \times x \times \frac{f_X(d_0)}{2(1-F_X(d_0))} \\ &\geq \bar{F}_X(d_0) \times x \times \frac{f_X(d_0)}{2(1-F_X(d_0))} > \gamma + k_0\end{aligned}$$

for all $x > 2(\gamma+k_0)/f_X(d_0)$. Now we know that the median of $Bin(n,p)$ is $\lfloor np \rfloor$ or $\lceil np \rceil$. For simplicity, we will use the bound here $\mathbb{P}(Bin(n,p) \geq np) \geq 1/2$ as it will be valid simply replacing np by $\lfloor np \rfloor$ and this will not alter any of our subsequent analysis. Therefore we have:

$$\begin{aligned}&\sum_{r=k_0}^{n} \frac{\mathbb{P}\left(\text{Bin}(r+\gamma, p_{x,n}) \geq \gamma + k_0\right)}{(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}(R_n = r) \\ &\geq \sum_{r=n\bar{F}(d_0)}^{n} \frac{\mathbb{P}\left(\text{Bin}(r+\gamma, p_{x,n}) \geq \gamma + k_0\right)}{(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}(R_n = r) \\ &\geq \sum_{r=n\bar{F}(d_0)}^{n} \frac{1}{2(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}(R_n = r) \\ &\geq \frac{1}{2(n+1)(n+2)\dots(n+\gamma)} \sum_{r=n\bar{F}(d_0)}^{n} \mathbb{P}(R_n = r) \\ &\geq \frac{1}{4(n+1)(n+2)\dots(n+\gamma)}.\end{aligned}$$

Going back to equation (B.19) we have:

$$\begin{aligned}&\mathbb{P}\left(n(\hat{d}-d_0) > x\right) \\ &\geq \frac{c_1 c_2 (p^*)^2}{\gamma} p_{x,n}^{-\gamma} \left[\sum_{r=k_0}^{n} \frac{\mathbb{P}\left(\text{Bin}(r+\gamma, p_{x,n}) \geq \gamma\right)}{(r+1)(r+2)\dots(r+\gamma)} \mathbb{P}(R_n = r) \right. \\ &\qquad \left. - p_{x,n}^{\gamma} \sum_{r=1}^{n} \frac{1}{(r+1)^{\gamma}} \mathbb{P}(R_n = r)\right] \\ &\geq \frac{c_1 c_2 (p^*)^2}{\gamma} p_{x,n}^{-\gamma} \left[\frac{1}{4(n+1)(n+2)\dots(n+\gamma)} - p_{x,n}^{\gamma} \frac{C(\bar{F}(d_0))^{-\gamma}}{n^{\gamma}}\right] \\ &\geq \frac{c_1 c_2 (p^*)^2}{\gamma} x^{\gamma} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \left[\frac{n^{\gamma}}{4(n+1)(n+2)\dots(n+\gamma)} \right. \\ &\qquad \left. - (np_{x,n})^{\gamma} \frac{C(\bar{F}(d_0))^{-\gamma}}{n^{\gamma}}\right] \\ &\geq \frac{c_1 c_2 (p^*)^2}{\gamma} x^{\gamma} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \left[\frac{1}{4(2^{\gamma})} - (np_{x,n})^{\gamma} \frac{C(\bar{F}(d_0))^{-\gamma}}{n^{\gamma}}\right]\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{c_1 c_2 (p^*)^2}{\gamma} x^\gamma \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \left[\frac{1}{4(2^\gamma)} - \left(x \times \frac{f_{X,\max}}{1-F_X(d_0)}\right)^\gamma \frac{C(\bar{F}(d_0))^{-\gamma}}{n^\gamma}\right] \\
&\geq \frac{c_1 c_2 (p^*)^2}{\gamma} x^\gamma \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \left[\frac{1}{4(2^\gamma)} - \left(\frac{\delta_2 f_{X,\max}}{C(1-F_X(d_0))^2}\right)^\gamma\right] \quad [\forall\ x \leq \delta_2 n] \\
&\geq \frac{c_1 c_2 (p^*)^2}{\gamma} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \times \frac{1}{2^{\gamma+3}} \times x^{-\gamma} \\
&= \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+3}} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \times x^{-\gamma}\,.
\end{aligned}
$$

where the last inequality is valid for small enough δ_2, i.e. we choose δ_2 which satisfies:

$$
\frac{1}{2^{\gamma+2}} - \left(\frac{\delta_2 f_{X,\max}}{C(1-F_X(d_0))^2}\right)^\gamma \geq \frac{1}{2^{\gamma+3}}\,.
$$

Therefore we have established that for any $2\gamma/f_X(d_0) \leq x \leq (\delta_1 \wedge \delta_2)n$:

$$
\mathbb{P}\left(n(\hat{d}-d_0) > x\right) \leq \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+3}} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \times x^{-\gamma}\,. \tag{B.20}
$$

The calculation for the negative x is similar. As introduced before, L_n denotes the number of $X_i's$ on the left of d_0 and $L_n \sim \text{Bin}(n, F(d_0))$. Given $L_n = l$, define $S_0 \equiv 0, S_1', \ldots, S_l'$ to be random walk on the left of origin and $S_0 = 0, S_1, S_2, \ldots, S_{n-l}$ on the right of origin. Given $L_n = l, N_{n,-}(x) = k'$, the event $n(\hat{d}-d_0) < -x$ is equivalent to the event that in a two sided random walk with l steps on the left and $n-l$ steps on the right, the minima occurs on the left and it occurs at one of the steps among $\{S_{k'+1}', S_{k'+2}', \ldots, S_l'\}$. Using the similar logic as above we obtain for any $2\gamma/f_X(d_0) \leq x \leq (\delta_1 \wedge \delta_2)n$:

$$
\mathbb{P}\left(n(\hat{d}-d_0) < -x\right) \geq \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+3}} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \times x^{-\gamma}\,. \tag{B.21}
$$

Finally, from equation (B.20) and (B.21) we conclude for any $2\gamma/f_X(d_0) \leq x \leq (\delta_1 \wedge \delta_2)n$:

$$
\mathbb{P}\left(\left|n(\hat{d}-d_0)\right| > x\right) \geq \frac{c_1 c_2 (p^*)^2}{\gamma 2^{\gamma+2}} \times \left(\frac{f_{X,\max}}{1-F_X(d_0)}\right)^{-\gamma} \times x^{-\gamma}\,.
$$

This completes the proof. □

B.7. Proof of Lemma A.6

We use the same notations as used in the proof of Lemma A.5. As ξ_i's are bounded by b with mean $\mu > 0$, by applying Hoeffding's inequality, we have for any $j \in \mathbb{N}$:

$$
\mathbb{P}\left(\bar{\xi}_j < -\mu\right) \leq e^{-\frac{j\mu^2}{4b^2}} := e^{-cj}\,,
$$

with $c = \mu^2/4b^2$. As in Lemma A.5, we start with establishing an upper bound on the tail on the minimizer of random walk. Let $\{S_j\}_{j=0,\ldots,n}$ denotes a n-step random walk and let Z_n denotes its minimizer supported on $\{0, 1, \ldots, n\}$. We then have:

$$\begin{aligned}
\mathbb{P}(Z_n > k) &= \sum_{j=k+1}^{n} \mathbb{P}(Z_n = j) \\
&= \sum_{j=k+1}^{n} \mathbb{P}(S_j < 0, \ldots, S_j < S_{j-1}, S_j < S_{j+1}, \ldots, S_j < S_n) \\
&\leq \sum_{j=k+1}^{n} \mathbb{P}(S_j < 0) \\
&= \sum_{j=k+1}^{n} \mathbb{P}\left(\bar{\xi}_j < -\mu\right) \\
&\leq \sum_{j=k+1}^{n} e^{-cj} = \frac{e^{-c(k+1)}}{1 - e^{-c}}\,.
\end{aligned}$$

Going back to distribution of $n(\hat{d} - d_0)$, as in the proof of Lemma A.5 we have:

$$\begin{aligned}
&\mathbb{P}\left(n(\hat{d} - d_0) > x\right) \\
&= \sum_{r=0}^{n}\sum_{k=0}^{r}\Big[\mathbb{P}\left(n(\hat{d} - d_0) > x \mid R_n = r, N_{n,+}(x) = k\right) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)\Big] \\
&= \sum_{r=0}^{n}\sum_{k=0}^{r}[\mathbb{P}(\text{argmin twosided RW} > k \mid R_n = r, N_{n,+}(x) = k) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \\
&= \sum_{r=0}^{n}\sum_{k=0}^{r}[\mathbb{P}(\text{argmin twosided RW} > k \mid R_n = r) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \qquad \text{(B.22)}
\end{aligned}$$

Upper bounding the argmin of two-sided random walk is relatively easier:

$$\begin{aligned}
&\mathbb{P}(\text{argmin twosided RW} > k \mid R_n = r) \\
&= \sum_{j=k+1}^{r} \mathbb{P}(\text{argmin twosided RW} = j \mid R_n = r) \\
&= \sum_{j=k+1}^{r} \mathbb{P}\left(S_j < \min_{0\leq i\leq j-1} S_i, S_j < \min_{j+1\leq i\leq r} S_i, S_j < \min_{1\leq i\leq n-r} S'_i\right) \\
&\leq \sum_{j=k+1}^{r} \mathbb{P}(S_j < 0) \leq \frac{e^{-c(k+1)}}{1 - e^{-c}}\,.
\end{aligned}$$

Using this bound in equation (B.22) we obtain for any $0 \le x \le n\delta_1$ (where δ_1 is same as defined in the proof of Lemma A.5, i.e. we choose $\delta_1 > 0$ such that $f_X(t) \ge f_X(d_0)/2$ for all $|t - d_0| \le \delta_1$):

$$\begin{aligned}
&\mathbb{P}\left(n(\hat{d} - d_0) > x\right) \\
&= \sum_{r=0}^{n}\sum_{k=0}^{r}[\mathbb{P}\left(\text{argmin twosided RW } > k \mid R_n = r\right) \\
&\qquad\qquad \times \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r)] \\
&\le \sum_{r=0}^{n}\sum_{k=0}^{r}\frac{e^{-c(k+1)}}{1-e^{-c}}\mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \\
&\le \frac{e^{-c}}{1-e^{-c}}\sum_{r=0}^{n}\sum_{k=0}^{r}e^{-ck}\ \mathbb{P}(N_{n,+}(x) = k \mid R_n = r)\mathbb{P}(R_n = r) \\
&= \frac{e^{-c}}{1-e^{-c}}\sum_{r=0}^{n}\left(1 - p_{n,x} + p_{n,x}e^{-c}\right)^r \mathbb{P}(R_n = r) \\
&= \frac{e^{-c}}{1-e^{-c}}\left(1 - \bar{F}(d_0) + \bar{F}(d_0)\left(1 - p_{n,x} + p_{n,x}e^{-c}\right)\right)^n \\
&= \frac{e^{-c}}{1-e^{-c}}\left(1 - \bar{F}(d_0)p_{n,x}(1-e^{-c})\right)^n \\
&\le \frac{e^{-c}}{1-e^{-c}}\left(1 - (1-e^{-c})\frac{xf_X(d_0)}{2n}\right)^n \\
&\le \frac{e^{-c}}{1-e^{-c}}e^{-x\frac{f_X(d_0)}{2}(1-e^{-c})}.
\end{aligned}$$

The calculation for $\mathbb{P}(n(\hat{d} - d_0) < -x)$ for $x > 0$ is similar and hence skipped for brevity. Therefore we obtain for $0 \le |x| \le n\delta_1$:

$$\mathbb{P}\left(\left|n(\hat{d} - d_0)\right| > x\right) \le \frac{2e^{-c}}{1-e^{-c}}e^{-x\frac{f_X(d_0)}{2}(1-e^{-c})}.$$

This completes the proof.

B.8. Proof of Proposition A.7

From proposition 5 of [17] we have:

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi f(X_i)\right\|_{\mathcal{F}} \le \mathbb{E}\left[\sum_{k=1}^{n}\left(\left|\eta_{(k)}\right| - \left|\eta_{(k+1)}\right|\right)\mathbb{E}\left\|\sum_{i=1}^{k}\epsilon_i f(X_i)\right\|_{\mathcal{F}}\right]$$

where $\left|\eta_{(1)}\right| \ge \left|\eta_{(2)}\right| \ge \ldots \left|\eta_{(n)}\right| \ge \left|\eta_{(n+1)}\right| = 0$ are the decreasing order statistics of $\{|\xi_i - \xi_i'|\}_{i=1}^{n}$ where $\{\xi_i'\}_{1\le i\le n}$ are i.i.d copy of $\{\xi_i\}_{1\le i\le n}$. Hence we have:

$$\mathbb{E}\left\|\sum_{i=1}^{n}\xi f(X_i)\right\|_{\mathcal{F}}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{k=1}^{n}\left(\left|\eta_{(k)}\right|-\left|\eta_{(k+1)}\right|\right)\mathbb{E}\left\|\sum_{i=1}^{k}\epsilon_i f(X_i)\right\|_{\mathcal{F}}\right]\\
&\leq \mathbb{E}\left[\sum_{k=1}^{n}\left(\left|\eta_{(k)}\right|-\left|\eta_{(k+1)}\right|\right)\left(\varphi_n(k)+b_n\right)\right]\\
&= \mathbb{E}\left[\sum_{i=1}^{n}\int_{\left|\eta_{(k+1)}\right|}^{\left|\eta_{(k)}\right|}\varphi_n(k)\ dt\right]+b_n\mathbb{E}\left[\left|\eta_{(1)}\right|\right]\\
&\leq \mathbb{E}\left[\int_0^{\infty}\varphi_n\left(\left|\{i:|\eta_i|\geq t\}\right|\right)\ dt\right]+b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i-\xi_i'|\right]\\
&\leq \int_0^{\infty}\varphi_n\left(\sum_{i=1}^{n}\mathbb{P}\left(|\xi_i-\xi_i'|>t\right)\right)+2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right] \qquad \text{[By Jensen's inequality]}\\
&\leq \int_0^{\infty}\varphi_n\left(\sum_{i=1}^{n}\left(\mathbb{P}\left(|\xi_i|>t/2\right)+\mathbb{P}\left(|\xi_i'|>t/2\right)\right)\right)+2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right]\\
&= \int_0^{\infty}\varphi_n\left(2\sum_{i=1}^{n}\mathbb{P}\left(|\xi_i|>t/2\right)\right)+2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right]\\
&= 2\int_0^{\infty}\varphi_n\left(2\sum_{i=1}^{n}\mathbb{P}\left(|\xi_i|>t\right)\right)+2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right]\\
&\leq 4\int_0^{\infty}\varphi_n\left(\sum_{i=1}^{n}\mathbb{P}\left(|\xi_i|>t\right)\right)+2b_n\mathbb{E}\left[\max_{1\leq i\leq n}|\xi_i|\right]
\end{aligned}
$$

where the last inequality follows from the fact that $\varphi_n(0)=0$ and φ_n concave which leads to $\varphi_n(2x)\leq 2\varphi_n(x)$.

B.9. Generalization of Theorem 2.4

In this subsection we prove a generalized version of Theorem 2.4 under the following tail assumption on the error distribution:

$$\mathbb{P}\left(|\xi|\geq t\right)\sim t^{-\gamma}$$

as mentioned in Remark 2.5. More precisely, we assume the following:

1. $\sup_{t\in[0,\infty)} t^{\gamma}\mathbb{P}\left(|\xi|\geq t\right)\leq C_U$,
2. $\sup_{t\in[t_0,\infty)} t^{\gamma}\mathbb{P}\left(|\xi|\geq t\right)\geq C_L(t_0)>0$ for all $t_0>0$ and $C_L(t_0)\downarrow 0$ as $t_0\downarrow 0$.

As before, we are also assuming ξ is symmetric. The proof of theorem is similar to Theorem 2.4. We will highlight the main differences here. Recall that the proof of Theorem 2.4 is based on four lemmas: Lemma A.1 - Lemma A.4. The proof of Lemma A.1 will not change, as it does not depend on the tail of the distribution of ξ. The proof of Lemma A.2 will be changed: The new version of that Lemma will be following:

Lemma B.1. *Suppose $\xi_1, \xi_2, \dots$ i.i.d. random variables with distribution symmetric around origin and satisfies the above mentioned tail conditions for some exponent γ. Define $X_i = \xi_i + \mu$ for some $\mu > 0$ and a random walk based on X_i's, i.e $S_n = \sum_{i=1}^n X_i$. Suppose M denotes the minimizer of the random walk on $\mathbb{Z}^+$. Then we have:*

$$\mathbb{P}(M \geq k) \geq \frac{c_1 c_2 p^*}{\gamma} \times \frac{1}{k^\gamma} := c_0 k^{-\gamma},$$

for all $k \geq k_0 := 1 \vee \lceil \mu^{-\gamma/(\gamma-1)} \rceil \vee \lceil 3C_U \rceil$, where:

1. $p^* = \mathbb{P}(S_i > 0 \ \ \forall \ \ i \in \mathbb{N}) = \mathbb{P}(M = 0)$,
2. $c_1 = \frac{C_L(\mu)}{(\mu+1)^\gamma}$,
3. $c_2 = \inf_{x \geq \lceil 3C_U \rceil} \left(1 - \frac{2C_U}{x}\right)^{x-1}$.

Proof. For any fixed $k \geq 1$ we have:

$$\begin{aligned}
\mathbb{P}(M \geq k) &= \sum_{j \geq k} \mathbb{P}(M = j) \\
&= \sum_{j \geq k} \mathbb{P}(S_i > S_j \ \forall \ 0 \leq i \leq j-1, S_i > S_j \ \forall \ i \geq j+1) \\
&= \mathbb{P}(S_1 > 0, S_2 > 0, \dots) \sum_{j \geq k} \mathbb{P}\left(\max_{1 \leq i \leq j} S_i < 0\right) \\
&= p^* \sum_{j \geq k} \mathbb{P}\left(\max_{1 \leq i \leq j} S_j < 0\right) \qquad [p^* = P(S_1 > 0, S_2 > 0, \dots)] \\
&\geq p^* \sum_{j \geq k} \frac{1}{j} P(S_j \leq 0) \qquad \text{(B.23)}
\end{aligned}$$

where the last inequality uses Lemma A.1. From the symmetry of the distribution of ξ we have:

$$\mathbb{P}(S_j \leq 0) = \mathbb{P}\left(\sum_{i=1}^j \xi_j \leq -j\mu\right) = \mathbb{P}\left(\sum_{i=1}^j \xi_j > j\mu\right).$$

Set $a_j = j^{1/\gamma}$. Define the event A_i as:

$$A_i = \{\xi_i > j\mu + (j-1)a_j, \ \xi_l \in [-a_j, j\mu) \ \forall \ 1 \leq l \neq i \leq j\}$$

Clearly $\{A_i\}'s$ are disjoint events and

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^j \xi_j > j\mu\right) &\geq \mathbb{P}\left(\cup_{i=1}^j A_i\right) \\
&= \sum_{i=1}^j \mathbb{P}(A_i)
\end{aligned}$$

$$= j\bar{F}\left(j\mu + (j-1)a_j\right)\left(F[-a_j, j\mu)\right)^{j-1}$$
$$= j\bar{F}\left(j\mu + (j-1)a_j\right)\left(1 - \bar{F}(a_j) - \bar{F}(j\mu)\right)^{j-1} \tag{B.24}$$

Next note that, $j\mu + (j-1)a_j \geq \mu$ for all $j \geq 1$. Therefore from equation (B.5) we have for all $j \geq 1$:

$$\bar{F}\left(j\mu + (j-1)a_j\right) \geq C_L(\mu)\left(j\mu + (j-1)a_j\right)^{-\gamma},$$

which further implies:

$$\begin{aligned}
j \times \bar{F}\left(j\mu + (j-1)a_j\right) &\geq C_L(\mu)\frac{j}{\left(j\mu + (j-1)a_j\right)^{\gamma}} \\
&= C_L(\mu)\frac{j}{\left(j\mu + (j-1)j^{1/\gamma}\right)^{\gamma}} \\
&= C_L(\mu)\frac{j}{j^{\gamma+1}\left(j^{-1/\gamma}\mu + \left(1 - \frac{1}{j}\right)\right)^{\gamma}} \\
&= \frac{C_L(\mu)}{j^{\gamma}}\ \frac{1}{\left(j^{-1/\gamma}\mu + \left(1 - \frac{1}{j}\right)\right)^{\gamma}} \\
&\geq \frac{C_L(\mu)}{j^{\gamma}}\ \frac{1}{(\mu+1)^{\gamma}} := \frac{c_1}{j^{\gamma}}.
\end{aligned}$$

Next observe that $(j\mu)^{\gamma} \geq j$ for all $j \geq 1$ if $\mu > 1$ or for all $j \geq \mu^{-\gamma/(\gamma-1)}$ if $\mu \leq 1$. Using this in equation (B.24) we have for all $j \geq 1 \vee \lceil \mu^{-\gamma/(\gamma-1)} \rceil \vee \lceil 3C_U \rceil$:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{j} \xi_j > j\mu\right) &\geq \frac{c_1}{j^{\gamma}}\left[\left(1 - \bar{F}(a_j) - \bar{F}(j\mu)\right)^{j-1}\right] \\
&\geq \frac{c_1}{j^{\gamma}}\left[\left(1 - C_U\left(a_j^{-\gamma} + (j\mu)^{-\gamma}\right)\right)^{j-1}\right] \\
&\geq \frac{c_1}{j^{\gamma}}\left[\left(1 - C_U\left(\frac{1}{j} + \frac{1}{(j\mu)^{\gamma}}\right)\right)^{j-1}\right] \\
&\geq \frac{c_1}{j^{\gamma}}\left[1 - \frac{2C_U}{j}\right]^{j-1} \qquad [\because j\mu \geq j] \\
&\geq \frac{c_1}{j^{\gamma}} \times \inf_{x \geq \lceil 3C_U \rceil}\left(1 - \frac{2C_U}{x}\right)^{x-1} \\
&:= \frac{c_1 c_2}{j^{\gamma}}
\end{aligned} \tag{B.25}$$

Using this in equation (B.23) we obtain:

$$P\left(M \geq k\right) \geq p^* \sum_{j \geq k} \frac{1}{j} P\left(S_j \leq 0\right)$$

$$
\begin{aligned}
&= p^* \sum_{j \geq k} \frac{1}{j} \left(\sum_{i=1}^{j} \xi_j > j\mu \right) \\
&\geq c_1 c_2 \times p^* \times \sum_{j \geq k} j^{-(\gamma+1)} \\
&\geq c_1 c_2 \times p^* \times \int_k^\infty x^{-(\gamma+1)} \, dx \\
&\geq c_1 c_2 \times p^* \times \frac{1}{\gamma k^\gamma} .
\end{aligned}
$$

This completes the proof of lower bound. □

The proofs of Lemma A.3, Lemma A.4 as well as the rest of the argument remain unchanged. This completes the proof of the generalized version of Theorem 2.4.

Appendix C: More simulations

In this section we present elaborate simulation results for various values of $k \in \{0.1, 0.5, 1, 2, 5, 10\}$ to complement the simulation in Section 4 of main paper, where simulations for ℓ_1 criterion ($k = 0$) and ℓ_2 criterion ($k = \infty$) are presented. Here also we conducted experiments for four different signal levels ($\mu \in \{0.1, 0.5, 1, 2\}$).

TABLE 9
Quantiles of asymptotic distribution with $\mu = 1, k = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	8.3536987880812	13.5927682886652	19.3609171467853	27.6663592062195	34.5579382045799
T_4	10.4788884372948	16.9486681760718	24.1027899555334	34.6517407531914	43.0321781340036
T_5	11.6978692249975	18.8922281852943	26.8795756073259	38.5230842513588	48.094967379471
T_6	12.4673379750485	20.1972893311096	28.8124291608022	41.3783772351445	51.6753216947574
T_{10}	14.0000855492778	22.7095710062678	32.5473987595816	46.5409929468191	58.0397031493322
T_{15}	14.6876914452795	23.8852566324027	34.0043421638412	48.6283574347637	60.4413437466999
Normal	15.9991878710632	26.1190617162937	37.3726341880332	53.5118731588664	66.5764957196666

TABLE 10
Quantiles of asymptotic distribution with $\mu = 1, k = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	8.32493407718205	13.5405276438179	19.3047074781908	27.5801370551298	34.3754080746644
T_4	10.3475765957395	16.6841313019627	23.7966698462179	34.0389358097088	42.2652055393164
T_5	11.4586343013648	18.5353104327393	26.3517929459007	37.7443111113224	46.7821424347838
T_6	12.1908646105817	19.721295780556	28.0266082386693	40.127409557103	49.8162604852832
T_{10}	13.5602636080797	21.9387825182329	31.2432135813082	44.5313945212685	55.5778874062068
T_{15}	14.239221889277	23.0289587553665	32.8163191650782	46.9208960085874	58.2658889881086
Normal	15.403873209345	24.8700967052205	35.3620562213101	50.7334136469591	63.0498352682136

TABLE 11
Quantiles of asymptotic distribution with $\mu = 1, k = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	8.62036825564386	14.0081909142818	19.9541673549796	28.5681359225452	35.524172881301
T_4	10.2990926166834	16.6203209925997	23.5039309935524	33.7036724352976	41.9678567620591
T_5	11.3276893731616	18.2469755527696	25.9557541469329	36.9843710045347	45.7497985547797
T_6	11.9135214024096	19.1984271121553	27.2774369123353	39.0181625769645	48.3123893082825
T_{10}	12.9998145790674	20.9518202195201	29.8029620277354	42.5882911278081	52.8099356616132
T_{15}	13.5241418961661	21.7777549923007	30.9538813220154	44.1251345928093	54.6843626784368
Normal	14.4150444145811	23.242553834748	33.0062982814038	47.335611181757	58.8616621356656

TABLE 12
Quantiles of asymptotic distribution with $\mu = 1, k = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	9.67668771770413	15.965777602917	22.9582079555854	32.8033850250549	40.944891675974
T_4	11.2541621802798	18.2141332407493	25.9528861199238	37.2196226574348	46.1626939615799
T_5	11.9529499925983	19.3573628081287	27.5154521525252	39.3285576062721	48.9183687173307
T_6	12.3009374477742	19.8598566733151	28.223223525972	40.3287408821538	49.9858703570862
T_{10}	13.013835672381	20.9861287337384	29.8783671274062	42.3566419472444	52.3859467595871
T_{15}	13.2017541203442	21.2354922665999	30.1620296088269	43.1616439990877	53.4856500437048
Normal	13.6328704116346	21.8011383314066	30.9358377012567	44.0613618414608	54.5118631397754

TABLE 13
Quantiles of asymptotic distribution with $\mu = 1, k = 5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	11.2272610783194	18.9816731937737	27.9037725374461	40.8689145728618	51.2147878961961
T_4	12.4095320315674	20.4980446391945	29.6237376742403	42.7732713429297	53.338109787115
T_5	12.8480963235841	21.0545979913169	30.2147632457062	43.4445264556501	53.6409869173766
T_6	13.0869935601188	21.2716762920823	30.281581405516	43.2958191460236	53.9145395675979
T_{10}	13.3285370910218	21.5350755916879	30.6627492515443	43.6970533424286	54.455095165621
T_{15}	13.4417138370947	21.6328579001887	30.7494906092105	43.724874562948	54.2177605737277
Normal	13.5544750188985	21.7470304441407	30.7819392318713	43.6843110525953	53.8235614179449

TABLE 14
Quantiles of asymptotic distribution with $\mu = 1, k = 10$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	11.6576668557376	20.0345314896088	29.7720008003837	44.4351920765485	55.8367073088179
T_4	12.6160178306568	20.869684341797	30.2766780259759	44.2424316961008	55.4957320207511
T_5	12.8851286373215	21.1813979715255	30.4191220480701	43.7844640867915	54.4545413832026
T_6	13.0882957766607	21.3174888193934	30.5260570501067	43.8159728436371	54.3507431127391
T_{10}	13.3530536086038	21.5328919607922	30.6639273685561	43.8038383653262	54.5911464163574
T_{15}	13.4434745691714	21.6310161127235	30.7537719788635	43.7295501273868	54.2240963710686
Normal	13.5544750188985	21.7470304441407	30.7819392318713	43.6843110525953	53.8235614179449

TABLE 15
Quantiles of asymptotic distribution with $\mu = 0.5, k = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	28.532672412875	46.8706471694321	67.6041756789722	97.4930857962577	121.961700301769
T_4	38.2819193124782	62.552045362741	89.7837169542353	129.145932269432	161.477146552752
T_5	44.5091949682024	72.6717097294959	104.036589987155	149.905989438607	187.409833195342
T_6	48.0901574047503	78.557305525532	112.843800505048	162.44555861772	201.816166904598
T_{10}	55.8897061168498	91.0822988352974	130.395778088274	187.299604893212	233.505708387893
T_{15}	59.306272284557	96.9668096977247	138.764385485374	198.429619389818	248.260823570918
Normal	65.8760109758601	107.818484039077	154.899858263991	223.055122634049	277.208421216919

TABLE 16
Quantiles of asymptotic distribution with $\mu = 0.5, k = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	27.3750759298099	44.7722420183833	64.4808160016137	93.4515684758776	116.875208288023
T_4	35.9675925299576	58.4302111069793	83.9852525535933	120.403426488691	149.756837388405
T_5	41.1319246413426	67.0824127387807	95.9937750691976	138.233634035094	172.249968046942
T_6	44.5722648865697	72.4761304495912	103.806356433929	148.70427114392	185.244152286579
T_{10}	50.9981570910429	82.917143865562	118.96541900344	170.15068173747	212.360238214554
T_{15}	54.0748000455022	88.2001830220234	126.317788452529	181.187230372619	224.580722486573
Normal	59.6053563347309	97.0407453665941	139.468447475864	201.296446721351	250.769440088992

TABLE 17
Quantiles of asymptotic distribution with $\mu = 0.5, k = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	28.9535753413177	47.2595220674185	67.9211327501115	97.6801405963068	122.198394032872
T_4	35.7791257830564	58.2347028499178	83.8219937346199	119.892190456548	149.39017387062
T_5	39.8634886345531	64.9443023944148	93.0527076134776	133.266607169204	165.659528999383
T_6	42.310755074163	68.9720353361756	98.6544306915552	141.491380452853	176.840137099276
T_{10}	47.110656565438	76.5991835723628	109.399889920527	156.7377007216	195.203793533811
T_{15}	49.4154656052842	80.6729964074326	115.043039837822	164.671507525883	204.419235348596
Normal	53.414978588033	86.9810421233553	124.638214842517	178.620948305988	222.563969726123

TABLE 18
Quantiles of asymptotic distribution with $\mu = 0.5, k = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	34.3788217466515	56.2891673228673	81.0009868993164	117.25518597126	146.481873302119
T_4	40.0528201827162	65.7183707599371	94.2308521020519	135.271780061382	167.889272979153
T_5	42.8936146374266	70.09292397618	100.361912179339	143.917734529639	179.088695832393
T_6	44.5217857123843	72.4655849071026	103.753547851116	148.785176765578	184.378255284364
T_{10}	47.1852585876096	76.7860842088517	109.796238145931	156.734067877825	194.152986950608
T_{15}	47.9408520452537	78.0650540743987	111.257576889225	159.010498171202	197.846312495383
Normal	49.5079582470008	80.5396888348225	115.188154986851	164.835820260727	204.931241705677

TABLE 19
Quantiles of asymptotic distribution with $\mu = 0.5, k = 5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	42.1220330307212	69.9598361767187	100.819915710102	146.102720595447	182.518546383375
T_4	45.9610219900353	75.4575607721563	108.831713066518	156.963383423823	195.394653994033
T_5	47.5979403918097	77.744069603015	111.308095761715	159.83139664115	198.68635726911
T_6	48.0353440259296	78.4929265937533	112.706744708086	161.978436676216	201.50718406659
T_{10}	48.7341339930379	79.2930128086947	113.404427574125	163.474496322987	202.546150712566
T_{15}	48.8095450833039	79.3327055867973	113.344432217353	162.621338606352	201.969247675431
Normal	48.8633478609145	79.3772893502566	113.306088433024	162.350810999429	202.504161072626

TABLE 20
Quantiles of asymptotic distribution with $\mu = 0.5, k = 10$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	45.224623872767	75.8003731357883	110.098123597561	158.793083024044	199.590840364891
T_4	47.2992606048524	77.9991235132896	112.372671292715	162.096966234746	203.457561666718
T_5	48.1957562353362	78.6848709763304	112.819291364713	162.453880135871	203.077921030432
T_6	48.3308500398421	78.9263327959001	113.55342323937	162.65488272043	203.19345674427
T_{10}	48.6314926249916	79.4221086617119	113.708496414293	163.530535124679	202.618434071806
T_{15}	48.7490404141274	79.2849943678726	113.438820919391	163.023810987999	202.407042473259
Normal	48.8633478609145	79.3772893502566	113.306088433024	162.350810999429	202.504161072626

TABLE 21
Quantiles of asymptotic distribution with $\mu = 0.1, k = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	691.585913275318	1116.80671170884	1551.34731818157	2060.33124492946	2297.91684597076
T_4	913.490253643434	1428.32138237031	1888.58345590786	2287.69275540892	2421.72559976642
T_5	1033.08671029929	1573.52184136739	2018.10390732023	2353.63360925043	2455.70422474729
T_6	1100.461189421	1655.67567035186	2082.16598588138	2379.49750165915	2467.72884756757
T_{10}	1226.93056184611	1792.60002082038	2184.5372973677	2418.6516997812	2491.82835831576
T_{15}	1272.06492206848	1832.83196545174	2212.64945013956	2427.91992104435	2497.35717079279
Normal	1361.47359182986	1923.19294206175	2269.04708309755	2446.66140280472	2508.42377451608

TABLE 22
Quantiles of asymptotic distribution with $\mu = 0.1, k = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	623.417244213784	1013.2800943545	1422.65670625147	1925.48034984088	2212.23204432912
T_4	817.27242623072	1292.72485278913	1741.37019765842	2196.23140048926	2378.24526951747
T_5	924.151195765237	1441.31438841089	1898.02088348303	2293.00276856422	2427.15444034888
T_6	988.749508377825	1515.6766410492	1966.56086889793	2332.38619298946	2442.80578417274
T_{10}	1098.63668719227	1649.78210170229	2082.25266840223	2381.96067963126	2471.48125740886
T_{15}	1151.17717379716	1709.43382876765	2122.86465647614	2395.42991168005	2478.2142350597
Normal	1231.93644908801	1794.26976272699	2187.6887314645	2420.59557912748	2493.24921520276

TABLE 23
Quantiles of asymptotic distribution with $\mu = 0.1, k = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	662.469748667074	1069.59641469967	1488.1893428298	1995.21729152715	2262.08257094163
T_4	809.85559864959	1280.564679245	1734.41158922081	2192.7478522624	2376.19658522409
T_5	889.329462387479	1390.84039160048	1853.67928413198	2273.44452384894	2415.50603357218
T_6	939.825901600757	1458.15510511643	1914.91329602495	2303.9339532509	2430.83310329224
T_{10}	1027.80326922858	1570.2907106716	2012.36886016401	2349.68903251833	2452.87904401359
T_{15}	1064.18152223244	1610.61586497468	2049.3864131254	2364.70555005751	2460.68987663165
Normal	1126.89287133307	1681.4575792726	2109.81898470445	2390.77800765456	2475.7154315963

TABLE 24
Quantiles of asymptotic distribution with $\mu = 0.1, k = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	789.810105581448	1253.77149241333	1712.43147720877	2171.40442804131	2365.44782764469
T_4	902.445389003925	1408.08154990501	1866.31675019539	2274.07881727991	2415.92629873592
T_5	955.408443515265	1480.12080018515	1927.25293264727	2310.12102208929	2434.30120102033
T_6	978.157326702502	1506.62737045575	1958.87019152396	2325.92566934938	2438.93246572574
T_{10}	1022.21051199468	1559.49171459449	2002.88197334422	2347.65937521505	2452.51581183309
T_{15}	1042.76433420544	1582.68094761371	2027.68775812454	2356.86928670928	2456.90382075004
Normal	1057.53733624309	1602.02718256746	2042.2215940587	2360.34876901845	2457.72725708744

TABLE 25
Quantiles of asymptotic distribution with $\mu = 0.1, k = 5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	957.276051207347	1479.73804966541	1938.14356167367	2319.27497095394	2439.70112898844
T_4	1017.22085603425	1553.17756140914	2004.122746336	2346.99445905095	2451.51996412323
T_5	1039.51740811484	1578.76768211778	2017.75474590447	2353.92775663355	2455.06432557587
T_6	1051.26536625162	1591.81114185629	2030.73341953253	2359.67710552763	2458.99170195467
T_{10}	1051.75513519349	1597.79865506523	2035.04944747562	2361.72996605265	2459.31613180123
T_{15}	1051.25308472394	1597.54447455308	2036.91506373065	2360.98420003756	2460.0376094193
Normal	1052.56465100535	1600.15783374558	2039.04719379446	2364.68752110372	2461.63108637474

TABLE 26
Quantiles of asymptotic distribution with $\mu = 0.1, k = 10$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1020.45712422267	1558.69123832535	2008.17176344346	2350.1143551918	2455.14935360474
T_4	1045.17076588147	1590.18369073296	2033.37858852914	2357.32960146002	2458.80007064842
T_5	1050.20319301615	1591.4427771334	2033.22781529606	2360.43552714178	2459.37158936661
T_6	1054.27325275541	1594.86346079493	2033.57107565913	2358.17367243726	2458.13877824713
T_{10}	1053.86050126658	1599.37633115038	2035.55626325774	2362.82927301231	2459.52821139428
T_{15}	1050.67107411527	1597.29022112193	2036.88972466631	2360.59177260682	2460.01170680134
Normal	1052.59926213318	1600.20677260164	2039.11470855037	2364.76147570037	2461.66518805169

TABLE 27
Quantiles of asymptotic distribution with $\mu = 2, k = 0.1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1.55427591730933	4.66329728625738	7.42349875218966	11.0539831418588	13.8497982832069
T_4	2.54447478495746	5.5560288545974	8.40747031617402	12.1779954637349	15.2398715214712
T_5	2.94182182123734	5.98112922689444	8.89651353275513	12.8663145685487	15.9074437543874
T_6	3.17370680665086	6.21384601113391	9.17803654033794	13.2855327616803	16.4779710510846
T_{10}	3.56632580724373	6.66844408727504	9.73996279978746	13.9221700001134	17.255837751101
T_{15}	3.70412367859872	6.80613024346193	9.93788472919249	14.1934767329633	17.5376574761972
Normal	3.95801814591642	7.1204641408512	10.331048755716	14.7277759318252	18.2281531850659

TABLE 28
Quantiles of asymptotic distribution with $\mu = 2, k = 0.5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1.58525426520645	4.70287937494892	7.43309614180873	11.0627373920382	13.8414193614277
T_4	2.52645189287737	5.52864069023889	8.36326543103941	12.2670827262151	15.2616532940261
T_5	2.90112257141207	5.92547388420335	8.86961063370793	12.8427354368825	15.9409529030012
T_6	3.14416432686496	6.17855580565703	9.15447016594051	13.1184280093813	16.3851692047376
T_{10}	3.51789575665061	6.59427235501635	9.65610554922482	13.8602174418393	17.1697151617247
T_{15}	3.62870965392304	6.71961868535226	9.83979744369365	14.1153561564069	17.5288354049719
Normal	3.87141746721187	7.00529689242012	10.1504135851152	14.4969219956899	17.9268002826894

TABLE 29
Quantiles of asymptotic distribution with $\mu = 2, k = 1$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1.56454142997244	4.69269752541939	7.44074198257859	11.0986984415153	13.8477690144218
T_4	2.5224111887514	5.5433839705459	8.41561535741103	12.2751341343976	15.2483299430723
T_5	2.88738725150022	5.92189370153633	8.87436889185247	12.8584581744366	15.9838012674835
T_6	3.09097803012668	6.13924522104549	9.10092522874141	13.1646658332534	16.36655599759
T_{10}	3.46026135619142	6.52784341934985	9.55722648164952	13.6973730031202	16.9801254211912
T_{15}	3.57516399900527	6.63944616656377	9.69556320407447	13.8832014179134	17.1606598058695
Normal	3.77564633142913	6.87011450797764	10.0348510106745	14.3186399200353	17.7781934348015

TABLE 30
Quantiles of asymptotic distribution with $\mu = 2, k = 2$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	1.74098381758006	4.96666468324748	7.84942023979826	11.7060098572495	14.7270442018646
T_4	2.5963730012743	5.70204336495744	8.64364221023573	12.6855503293643	15.776343897399
T_5	2.95422548241382	6.04290885413885	9.07342365078278	13.1366020086245	16.3310970145165
T_6	3.15175835835109	6.23709313589106	9.25018421651132	13.3698992407219	16.6389753242163
T_{10}	3.42642481518157	6.48915398161272	9.52789378863859	13.6490634800389	16.8578922529491
T_{15}	3.55829926352428	6.60756704454375	9.60803550932904	13.7149496871267	16.9792927148309
Normal	3.73251802650301	6.77197360515593	9.7939009218535	13.9170081065767	17.1809967262489

TABLE 31
Quantiles of asymptotic distribution with $\mu = 2, k = 5$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	2.1126833976562	5.58062708411904	8.93192713966539	13.6521891674556	17.3472144354772
T_4	2.84957480202987	6.10157175982755	9.34800959375619	13.9547641402722	17.547950645855
T_5	3.10733663833669	6.32679322717114	9.52031208078865	13.967142616375	17.4126757056361
T_6	3.27307545220797	6.40815227181175	9.59182907385158	13.9425141513821	17.3794427574096
T_{10}	3.51639124107138	6.623312824866	9.73507323840116	13.9896348468686	17.3521699189524
T_{15}	3.581835769269	6.64932878743327	9.73100885700681	14.0103365846951	17.3162731750811
Normal	3.70041133579528	6.72991527356395	9.75483649073316	13.8292349748401	17.0074984099981

TABLE 32
Quantiles of asymptotic distribution with $\mu = 2, k = 10$

Distributions	90%	95%	97.50%	99%	99.50%
T_3	2.20826097242346	5.72123725896405	9.22079229374703	14.432918779651	18.7628631277345
T_4	2.87197941748629	6.1875014516893	9.53836253474882	14.3273558002239	18.2739382343264
T_5	3.13065012208848	6.37445273469656	9.61601462012811	14.1326437904448	17.7098321783359
T_6	3.30976886070275	6.48543857844941	9.68472717330602	14.083300810493	17.6512026395697
T_{10}	3.52374824364879	6.61361740990017	9.70135285982052	13.9793650927543	17.3477739282867
T_{15}	3.581835769269	6.64978368371662	9.7316939616228	14.0100410832258	17.3162731750811
Normal	3.70041133579528	6.72991527356395	9.75483649073316	13.8292349748401	17.0074984099981

References

[1] Felix Abramovich and Vadim Grinshtein. High-dimensional classification by sparse logistic regression. *arXiv preprint arXiv:*1706.08344, 2017. MR3951383

[2] José A Adell and P Jodrá. The median of the poisson distribution. *Metrika*, 61(3):337–346, 2005. MR2230380

[3] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999. MR1679028

[4] Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

[5] S Basso, M Schirmer, and G Botter. On the emergence of heavy-tailed streamflow distributions. *Advances in Water Resources*, 82:98–105, 2015.

[6] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

[7] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013. MR0233396

[8] Lucien Birgé. A new lower bound for multiple hypothesis testing. *IEEE transactions on information theory*, 51(4):1611–1615, 2005. MR2241522

[9] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013. MR3185193

[10] Brendan O Bradley and Murad S Taqqu. Financial risk and heavy tails. In *Handbook of heavy tailed distributions in finance*, pages 35–103. Elsevier, 2003.

[11] Francisco Cribari-Neto, Nancy Lopes Garcia, and Klaus LP Vasconcellos. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2):269–277, 2000.

[12] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009. MR4158199

[13] Herbert Edelsbrunner. *Algorithms in combinatorial geometry*, volume 10. Springer Science & Business Media, 2012. MR0904271

[14] Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[15] Ailin Fan, Rui Song, and Wenbin Lu. Change-plane analysis for subgroup detection and sample size calculation. *Journal of the American Statistical Association*, 112(518):769–778, 2017. MR3671769

[16] William Feller. An introduction to probability theory and its applications. *1957*. MR0088081

[17] Qiyang Han, Jon A Wellner, et al. Convergence rates of least squares regression estimators with heavy-tailed errors. *Annals of Statistics*, 47(4):2286–2319, 2019. MR3953452

[18] Felix Hernandez-Campos, JS Marron, Gennady Samorodnitsky, and F Donelson Smith. Variable heavy tails in internet traffic. *Performance Evaluation*, 58(2-3):261–284, 2004.

[19] Joel L Horowitz. A smoothed maximum score estimator for the binary response model. *Econometrica: journal of the Econometric Society*, pages 505–531, 1992. MR1162997

[20] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[21] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004. MR0606374

[22] Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax theory of image reconstruction*, volume 82. Springer Science & Business Media, 2012. MR1226450

[23] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007. MR2724368

[24] Michael R Kosorok and Rui Song. Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics*, 35(3):957–989, 2007. MR2341694

[25] M.R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008. MR2724368

[26] Yan Lan, Moulinath Banerjee, and George Michailidis. Change-point estimation under adaptive sampling. *The Annals of Statistics*, 37(4):1752–1791, 2009. MR2533471

[27] Jialiang Li, Yaguang Li, and Baisuo Jin. Multi-threshold change plane model: Estimation theory and applications in subgroup identification. *arXiv preprint arXiv:*1808.00647, 2018.

[28] Clive R Loader. Change point estimation using nonparametric regression.

The Annals of Statistics, 24(4):1667–1678, 1996. MR1416655

[29] Pascal Massart. Concentration inequalities and model selection. MR2319879

[30] Pascal Massart, Élodie Nédélec, et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. MR2291502

[31] Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov. Non-standard asymptotics in high dimensions: Manski's maximum score estimator revisited. *arXiv preprint arXiv:*1903.10063, 2019.

[32] Debarghya Mukherjee, Moulinath Banerjee, and Ya'acov Ritov. Asymptotic normality of a linear threshold estimator in fixed dimension with near-optimal rate. *arXiv preprint arXiv:*2001.06955, 2020.

[33] Marc Raimondo. Minimax estimation of sharp change points. *Annals of statistics*, pages 1379–1397, 1998. MR1647673

[34] Patricia Reynaud-Bouret. Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities. *Probability Theory and Related Fields*, 126(1):103–153, 2003. MR1981635

[35] Bodhisattva Sen. A gentle introduction to empirical process theory and applications. 2018.

[36] M.H. Seo and O. Linton. A smoothed least squares estimator for threshold regression models. *Journal of Econometrics*, 141(2):704–735, 2007. MR2413485

[37] Noa Slater, Yoram Louzoun, Loren Gragert, Martin Maiers, Ansu Chatterjee, and Mark Albrecht. Power laws for heavy-tailed distributions: Modeling allele and haplotype diversity for the national marrow donor program. *PLoS Comput Biol*, 11(4):e1004204, 2015.

[38] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996. MR1385671

[39] Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

[40] Susan Wei and M.R. Kosorok. The Cox proportional hazards model with change plane. *Submitted*, 2015.

[41] Martin L Weitzman. Fat-tailed uncertainty in the economics of catastrophic climate change. *Review of Environmental Economics and Policy*, 5(2):275–292, 2011.

[42] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997. MR1462963