Maximum likelihood estimation of a unimodal probability mass function

Fadoua Balabdaoui and Hanna Jankowski

October 5, 2015

Abstract

We develop an estimation procedure of a discrete probability mass function (pmf) with unknown support. We derive the maximum likelihood estimator of this pmf under the mild and natural shape-constraint of unimodality. Shape-constrained estimation is a powerful and robust technique, which additionally provides smoothing of the empirical distribution yielding thereby gains in efficiency. We show that our unimodal estimator is consistent when the model is specified, and that it converges to the best projection of the true pmf on the unimodal class under model misspecification. Furthermore, we derive the limiting distribution of the the estimator, and use the obtained result to build asymptotic confidence bands for the unknown pmf when the latter is unimodal. We illustrate our approach using time-to-onset data of the Ebola virus during the 1976 outbreak in the former republic of Zaire.

1 Introduction

Discrete or discretized data show up in many practical instances, see Harlan et al. (2014); Chowell et al. (2013, 2009); Laskowski et al. (2011); Breman and Johnson (2014). If computing the empirical distribution requires no assumptions on the unknown law, gains in efficiency can be made by imposing additional constraints. Such a constraint is unimodality, which is a natural and mild assumption in many real statistical applications.

Nonparametric estimation of a unimodal density has been treated in many research papers. In case the mode is known, the problem boils down to fitting the well-known Grenander estimator (Grenander, 1956). However, as noted by Birgé (1997), it is unrealistic in practice to assume that the location of the mode is known. The main consequence of not making such an assumption is that the maximum likelihood estimator (MLE) fails to exist. To address this problem, several estimators have been proposed, see Wegman (1968, 1969); Prakasa Rao (1969); Wegman (1970a,b); Reiss (1973, 1976) in which the Grenander estimator has been additionally constrained.

More recent work appears in Birgé (1997), where the proposed estimator is chosen among all the possible unimodal Grenander estimators as the one with cumulative distribution function closest to the empirical distribution. Durot et al. (2013) consider the estimation of a discrete convex distribution, using the least squares criterion. Recently, Dümbgen and Rufibach (2009); Cule et al. (2010) proposed to use the maximum likelihood estimator of a log-concave density in lieu of the unimodal assumption, partially due to the inherent problems faced when estimating a unimodal density using maximum likelihood. As opposed to the continuous setting, existence of the unimodal MLE when the data are discrete is guaranteed, even when the mode is unknown. On the other hand, uniqueness is not always true, but this problem is rather marginal, as a rule for selecting from among the *finite* options is immediate, making our estimator fully automatic and easy to compute. Furthermore, if the pmf is not unimodal, the MLE is still consistent, in the sense that it approaches the best unimodal pmf among a finite number of choices. Further details of this behavior are provided in Section 4.

In the recent work of Balabdaoui et al. (2013), the discrete MLE under the constraint of log-concavity was studied. One important consequence of this work is that we can evaluate the loss when data exhibit unimodality but at the same time log-concavity would not be a valid assumption. The unimodal MLE seems to be a more natural estimator to consider when additional features of the true distribution besides unimodality are lacking or hard to obtain. On the other hand, it is expected the log-concave MLE to be more efficient than the unimodal one in case log-concavity is a correct assumption about the model. This is studied via simulations in Section 3. Although restricted to discrete distributions, our results may be interesting to those studying the continuous setting as well.

The manuscript is organized as follows. In Section 2, we provide the technical details required to define and compute the MLE of a discrete unimodal distribution. In our set-up, the support is assumed to be unknown, and is also estimated empirically from the data. In Section 3, we consider the finite sample size behavior of our estimator via simulations. As mentioned previously, we compare here our estimator with the discrete log-concave MLE, but also we assess the loss of efficiency when the support is unknown and must be estimated from the data. Sections 4 and 5 we establish consistency and global asymptotic theory for the estimator. One of our key contributions is the application of these to develop global confidence bands for a unimodal pmf, see Section 6. Finally, we illustrate the estimator on a data set for the 1976 Ebola outbreak in Zaire; see Section 7. The data clearly shows a drastic difference in the time from infection to onset of symptoms depending on the type of infection: whether the individual was infected from person-to-person contact or from injection with an unsterilized needle. `R` (R Core Team, 2014) code for this analysis (along with all simulations) is available online at `www.math.yorku.ca/~hkj/Research/`. All proofs and additional details are left to the Appendices.

2 Maximum likelihood estimation

2.1 Discrete unimodal distributions

In this work, we consider estimation of a unimodal pmf of a discrete real-valued random variable. We denote the support of such a pmf as $S = \{s_i\}_{i \in K}$, where K is a subset of $\mathbb{Z}$. Without loss of generality, we take $s_i \in \mathbb{R}$ for all $i \in K$, and we assume that $s_i < s_{i+1}$.

We say that a pmf p is unimodal if there exists an integer m such that

$$\begin{aligned} p(s_i) &\geq p(s_{i+1}), \quad \text{for all} \quad i \geq m, \quad \text{and} \\ p(s_{i-1}) &\leq p(s_i), \qquad \text{for all} \quad i \leq m. \end{aligned} \tag{2.1}$$

The element s_m is thus a mode of the pmf p, but is not necessarily unique. In general, we can define the *modal region*, denoted here by $\mathcal{M}$, as

$$\mathcal{M} = \{s_\kappa \in S : \ p \text{ satisfies (2.1) at } m = \kappa\} \tag{2.2}$$

$\mathcal{M}$ is necessarily a finite set and we have that $p(s) = p(s')$ for all $s, s' \in \mathcal{M}$. Next, let $\mathcal{U}^1(S)$ denote the space of unimodal pmfs with the same fixed support S. For the purpose of estimating such a p, it is most convenient to decompose the space of unimodal pmfs as

$$\mathcal{U}^1(S) = \bigcup_{\kappa \in K} \mathcal{U}^1|_\kappa(S), \tag{2.3}$$

where $\mathcal{U}^1|_\kappa(S)$ is the space of pmfs which are increasing on $\{s_i : i \leq \kappa - 1\}$ and decreasing on $\{s_i : i \geq \kappa\}$. Note that a pmf in $\mathcal{U}^1|_\kappa(S)$ is unimodal either at $s_{\kappa-1}$ or s_κ depending on the order of its values at these points. It may seem, at first, that it would be more natural to decompose $\mathcal{U}^1(S)$ into the spaces of pmfs that are unimodal at κ. However, it turned out that the decomposition (2.3) is much more convenient. In addition, the MLE will always "decide" between these two possibilities by choosing the one that yields the largest value of the likelihood. Note also that if $\kappa = \min K$, then $\mathcal{U}^1|_\kappa(S)$ is simply the space of non-increasing pmfs on S. Notably, each space $\mathcal{U}^1|_\kappa(S)$ is convex, whereas $\mathcal{U}^1(S)$ is not.

Known as Khintchine's Theorem, a density with respect to Lebesgue measure is unimodal if and only if it can be written as a mixture of uniform densities, see for example Olshen and Savage (1970). Hence, it is expected that such a representation exists also in the discrete setting.

Proposition 2.1. *A pmf p satisfies $p \in \mathcal{U}^1|_\kappa(S)$ if and only if*

$$p(s_i) = \sum_{j \geq 0} \frac{1_{i \in \{\kappa, \ldots, \kappa+j\}}}{j+1} q(s_j) + \sum_{j \leq -1} \frac{1_{i \in \{\kappa+j, \ldots, \kappa-1\}}}{|j|} q(s_j), \tag{2.4}$$

for some pmf q with support S.

A proof of Proposition 2.1 can be found in the Appendix. Using (2.3), a unimodal $p \in \mathcal{U}^1(S)$ admits a representation (2.4) for some $\kappa \in K$.

Remark 2.2. *Suppose that K is finite, and write $K = \{0, 1, \ldots, k\}$. Then $\mathcal{U}^1|_0(S) \subset \mathcal{U}^1|_1(S)$.*

2.1.1 Relationship with unimodal densities

Given a probability mass function p with support on $S = \mathbb{Z}$, one can define a density function f on $\mathbb{R}$ by

$$f(x) \quad = \quad p(z) \qquad \text{for } x \in (z-1, z], \ z \in \mathbb{Z}.$$

The mass function p is unimodal iff the (piecewise constant) density f is unimodal.

Given a general unimodal density with support on $\mathbb{R}$, one can also define a unimodal pmf p via

$$p(z) \quad = \quad \int_{z-1}^{z} f(x) dx, \quad z \in \mathbb{Z}.$$

Here, the choice of the discretization on $[z-1, z)$ is arbitrary. Indeed, any choice of $a \in \mathbb{R}$, with $[z+a-1, z+a)$ is possible. One could also consider intervals of length other than one, as long as the length is fixed.

In this sense, discrete distributions provide a useful way to analyze data which has been "discretized" in such a manner. One such example is considered in Section 7. This relationship with unimodal densities is particularly noteworthy, since, although the MLE of a unimodal density does not exist, the MLE of its discretized version does.

2.1.2 When the true support is unknown

The discussion above relating unimodal densities and pmfs implies that one natural assumption on the support S is that it is a connected subset of $a + \delta\mathbb{Z}$, for some $a \in \mathbb{R}$ and $\delta > 0$. However, we believe that in certain instances additional generality may be required. For this reason, the only assumption we make about the support S is that it is an ordered subset of $\mathbb{R}$. This assumption provides additional flexibility to our approach: unimodality of a pmf is preserved under scalar transformations (if the pmf of a random variable X is unimodal, then so is the pmf of aX), and under removal of elements of the support.

In order to reflect this flexibility in our estimation approach, we do not assume that the true support is known a priori. Instead, we estimate both the unimodal pmf and its support from the collected observations. However, if the true support is known a priori, then it is expected that more efficiency would be gained by including this information to the estimation procedure. Some simulations studying this are given in Section 3.1.

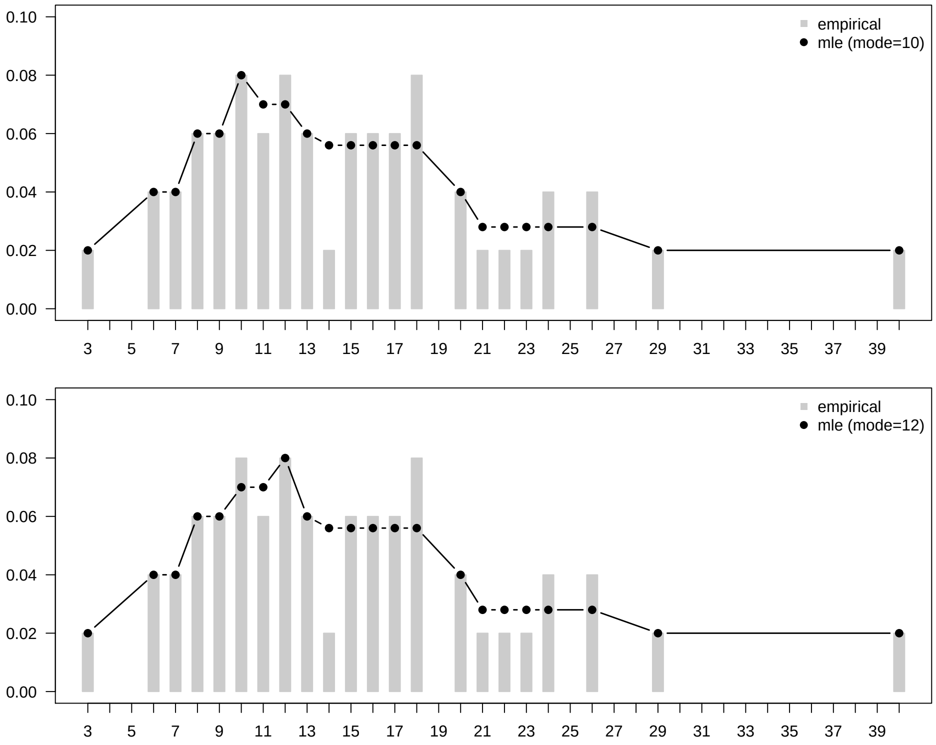

Figure 1: The same empirical observations (shown in grey) yield two different solutions maximizing the likelihood.

Notably, our consistency and asymptotic results developed later apply to both versions of the MLE (either when the support is known or when it is unknown). All theoretical results are proved and stated for the unknown support version; the proofs are only simplified when the support is known.

2.2 The unimodal maximum likelihood estimator

Let $X_1, \cdots, X_n$ be n independent observations from a discrete pmf p_0. Also, let $\overline{p}_n(z) = n^{-1}\sum_{j=1}^{n} \mathbb{I}_{\{z\}}(X_i)$ and $\mathbb{F}_n(z) = n^{-1}\sum_{j=1}^{n} \mathbb{I}_{\{(-\infty,z]\}}(X_i)$ denote the empirical pmf and the associated empirical cumulative distribution function (cdf), respectively. Finally, let S_n denote the observed support of $\overline{p}_n$, that is $S_n = \{z_0, \ldots, z_{J-1}\}$ is the set of distinct values in the sample $\{X_1, \ldots, X_n\}$. We assume in our notation that $z_0 < z_1 < \ldots < z_{J-1}$.

2.2.1 Definition

Recall that we do not assume that the support S is known. Thus, we define the maximum likelihood estimator (MLE) as

$$\widehat{p}_n \quad = \quad \operatorname{argmax}_{p \in \mathcal{U}^1(S_n)} \; L_n(p),$$

where the log-likelihood is given by

$$L_n(p) \quad = \quad \int \log p(z)\, d\mathbb{F}_n(z) \quad = \quad \sum_{j=0}^{J-1} \log(p(z_j))\, \overline{p}_n(z_j).$$

This maximization is done in two steps: (1) we maximize L_n over the space $\mathcal{U}^1|_\kappa(S_n)$ for each κ, and (2) we find $\widehat{\kappa}_n$ and the corresponding estimator at which the overall maximum is attained.

2.2.2 The shape operators iso, anti, and uni

To describe and compute the MLE, we find it convenient to first define several shape operators. For any $z \in \mathbb{R}^d$, we denote $z_{s:t}$ the sub-vector $(z_s, \ldots, z_t)$ where $1 \le s \le t \le d$. Consider the following sets of constrained vectors

$$\begin{aligned} \mathcal{I}_d &= \left\{u = (u_1, \ldots, u_d) \in \mathbb{R}^d : u_1 \le \cdots \le u_d\right\} \\ \mathcal{D}_d &= \left\{w = (w_1, \ldots, w_d) \in \mathbb{R}^d : w_1 \ge \cdots \ge w_d\right\}. \end{aligned}$$

Also, for $\kappa \in \{1, \ldots, d\}$, let

$$\begin{aligned} \mathcal{U}_d|_\kappa &= \left\{z = (z_1, \ldots, z_d) \in \mathbb{R}^d : z_{1:(\kappa-1)} \in \mathcal{I}_{\kappa-1} \text{ and } z_{\kappa:d} \in \mathcal{D}_{d-\kappa+1}\right\}, \quad \text{and} \\ \mathcal{U}_d &= \cup_{\kappa=1}^d \mathcal{U}_d|_\kappa. \end{aligned}$$

Lastly, we denote the ℓ_2 distance by $\|v-u\|_2^2 = \sum_{j=1}^d (v_j - u_j)^2$.

We can now define the first two operators $\mathrm{iso} : \mathbb{R}^d \to \mathcal{I}_d$ and $\mathrm{anti} : \mathbb{R}^d \to \mathcal{D}_d$ as

$$\begin{aligned} \mathrm{iso}[v] &= \mathrm{argmin}_{u \in \mathcal{I}_d} \|v - u\|_2 \\ \mathrm{anti}[v] &= \mathrm{argmin}_{w \in \mathcal{D}_d} \|w - u\|_2. \end{aligned}$$

In other words, $\mathrm{iso}[v]$ and $\mathrm{anti}[v] = -\,\mathrm{iso}[-v]$ are the well-known least squares projections of v on the spaces $\mathcal{I}_d$ and $\mathcal{D}_d$ respectively; cf. Barlow et al. (1972); Sen and Meyer (2013). Note also that the operator anti is the same as the gren operator discussed in Jankowski and Wellner (2009); Jankowski (2014).

Finally, for $\kappa \in \{1, \ldots, d\}$, define the operators $\mathrm{uni}_\kappa : \mathbb{R}^d \to \mathcal{U}_d|_\kappa$ and $\mathrm{uni} : \mathbb{R}^d \to \mathcal{U}_d$ as

$$\begin{aligned} \mathrm{uni}_\kappa[v] &= (\mathrm{iso}[v_{1:(\kappa-1)}], \mathrm{anti}[v_{\kappa:d}]) \quad = \quad \mathrm{argmin}_{u \in \mathcal{U}_d|_\kappa} \|v - u\|_2, \\ \mathrm{uni}[v] &= \mathrm{argmin}_{u \in \mathcal{U}_d} \|v - u\|_2. \end{aligned}$$

Note that, as before, we have that

$$\mathrm{uni}[v] \quad = \quad \mathrm{uni}_{\kappa=\widetilde{\kappa}}[v], \quad \text{where } \widetilde{\kappa} \quad \in \quad \mathrm{argmin}_\kappa \|v - \mathrm{uni}_\kappa[v]\|_2.$$

The operators iso and anti are unique. However, the operator uni may yield more than one solution, much like the operator yielding the MLE. Properties of these operators are discussed in detail in Appendix C.3.

2.2.3 Existence and characterization of the MLE

Using these operators, we may now state some facts about the MLE.

Proposition 2.3. *The restricted MLE $\widehat{p}_n|_\kappa$ exists and is unique. Furthermore, it is characterized by*

$$\widehat{p}_n|_\kappa \quad = \quad \mathrm{uni}_\kappa[\overline{p}_n].$$

The (unrestricted) unimodal MLE $\widehat{p}_n$ exists, but is not necessarily unique. For $\{\widehat{\kappa}_n\} = \mathrm{argmax}_{1\le\kappa\le J-1} L_n(\widehat{p}_n|_\kappa)$, the (finite) collection of solutions to the maximization problem, $\{\widehat{p}_n\}$, is characterized as

$$\{\widehat{p}_n\} \quad = \quad \{\widehat{p}_n|_\kappa ; \kappa \in \{\widehat{\kappa}_n\}\}. \tag{2.5}$$

Note that the MLE is not defined in terms of the operator uni, however, the operator does show up in its limiting distribution.

Remark 2.4. *One of the key conclusions of Proposition 2.3 is that the size of the set $\{\widehat{p}_n\}$ may be greater than one (see, for example, Figure 1). To overcome the computational difficulties that would result from this non-uniqueness, we simply take the MLE to be equal to the maximizer with the smallest mode. That is, let $\widehat{\kappa}_n$ denote the smallest integer κ such that*

$$L_n(\widehat{p}_n|_\kappa) \quad = \quad \max_{1\le l\le J-1} L_n(\widehat{p}_n|_l).$$

Then $\widehat{p}_n = \widehat{p}_n|_{\widehat{\kappa}_n}$. Note the slight abuse of notation: we denote $\widehat{\kappa}_n$ as the smallest element of $\{\widehat{\kappa}_n\}$. Also, note that in order to find $\widehat{\kappa}_n$ we can search only over $1 \le \kappa \le J-1$ using Remark 2.2.

The characterization in (2.5) along with our convention provides a straightforward way to compute $\widehat{p}_n$. Namely, we first find the restricted MLE $\widehat{p}_n|_\kappa$ as the right slopes of the greatest convex minorant of $\{(0,0),(z_j,\mathbb{F}_n(z_j),\ 0\le j\le \kappa-1\}$ and the left slopes of the least concave majorant of $\{(0,0),(z_j,\mathbb{F}_n(z_j)),\ \kappa\le j\le J-1\}$. The MLE $\widehat{p}_n$ will be then taken to be equal to $\widehat{p}_n|_\kappa$ which maximizes the overall likelihood for the smallest integer κ. Proposition 2.3 follows immediately from the more general result of Theorem 4.2 as well as the subsequent Lemma C.3.

3 Finite sample performance of the MLE

Here we compare three maximum likelihood estimators for small and medium samples sizes. The three estimators are

(1) the MLE under no assumption on the pmf; i.e., the empirical MLE,

(2) the MLE assuming the pmf is unimodal ($\widehat{p}_n$ as defined in this work),

(3) the log-concave MLE assuming the pmf is log-concave. Theoretical and computational aspects of this estimator have been studied in Balabdaoui et al. (2013).

In our simulations, we consider six different distributions:

- The negative binomial distribution with parameters $r = 6, p = 0.3$. This is a distribution is both strictly unimodal and strictly log-concave.
- The double logarithmic distribution with $S = \mathbb{Z}$, which we define as

$$p(z) \quad = \quad \begin{cases} \frac{p^{|z|}}{2|z|(p-\log(1-p))} & z \leq -1 \\ \frac{p}{p-\log(1-p)} & z = 0, \\ \frac{p^{z}}{2z(p-\log(1-p))} & z \geq 1. \end{cases} \tag{3.6}$$

  This distribution is strictly unimodal but not log-concave. In the simulations, we take $p = 0.9$.

- The uniform pmf with $S = \{0, \ldots, 9\}$. This is an example of a pmf which is neither strictly unimodal nor strictly log-concave.
- The mixture of uniform distributions with support on $\{0, \ldots, 49\}$, with pmf given by taking $S = \mathbb{Z}, \kappa = 0$ and

$$q(z) \quad = \quad \begin{cases} 1/3 & z = 9, 39, 49 \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

  in decomposition (2.1). This distribution is unimodal (though not strictly unimodal), and is not log-concave.

- The Poisson with rate $\lambda = 2$, a strictly log-concave and unimodal distribution.
- A mixture of Poisson distributions: Letting p_λ denote the pmf of a Poisson distribution with rate λ, then the mixture we consider is given by $(1/4) \cdot p_1(\cdot) + (1/8) \cdot p_3(\cdot) + (5/8) \cdot p_8(\cdot)$. This distribution is (strictly) bimodal, and is therefore neither unimodal nor log-concave.

Table 1: Properties of distributions considered.

	unimodal	log-concave	finite support
negative binomial	yes (strict)	yes (strict)	no
double logarithmic	yes (strict)	no	no
uniform	yes	yes	yes
uniform mixture	yes	no	yes
Poisson	yes (strict)	yes (strict)	no
Poisson mixture	no	no	no

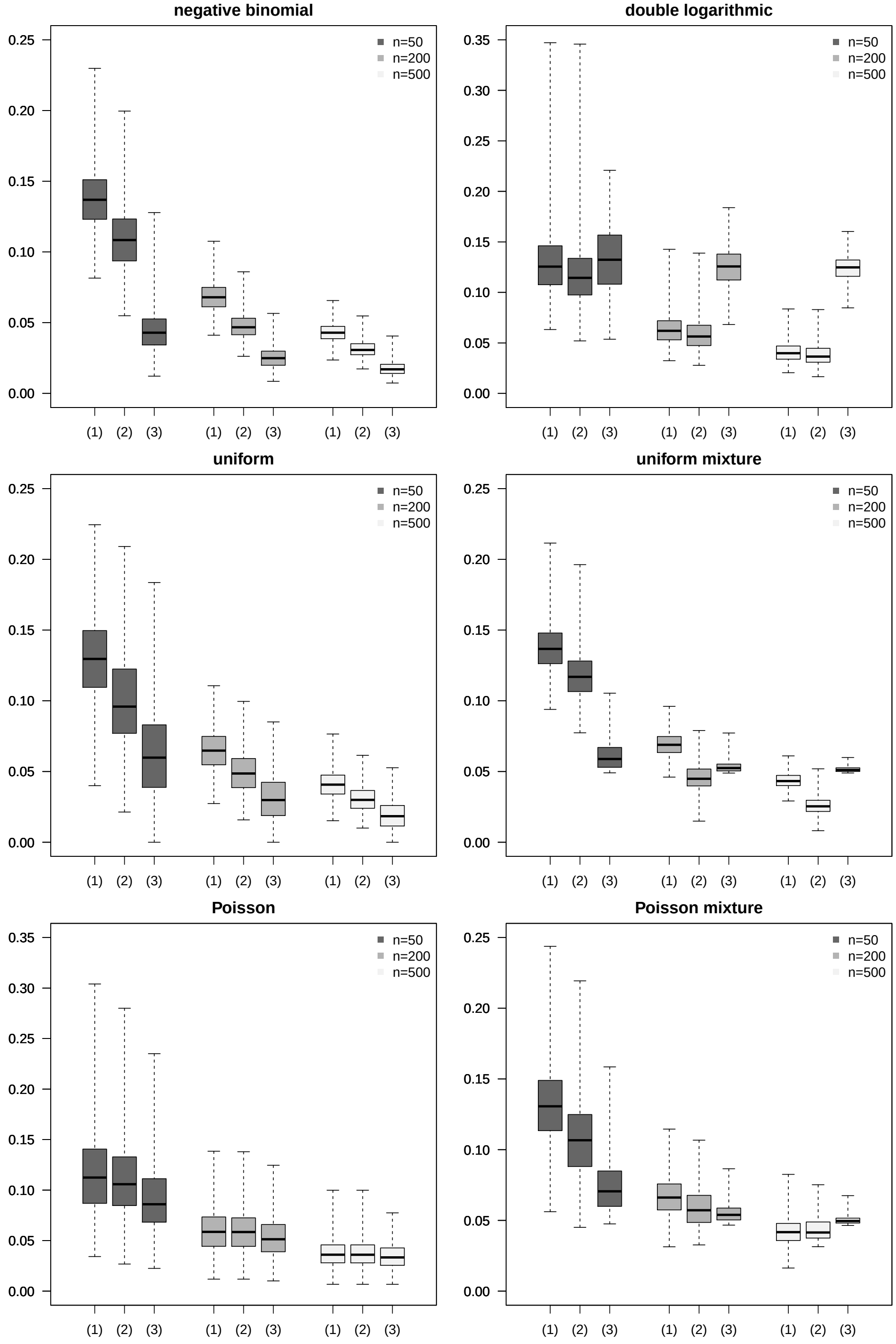

Figure 2: Boxplots of the ℓ_2 distance of the estimated pmf from the true pmf under each of three estimators: the empirical MLE (1), the unimodal MLE (2), the log-concave MLE (3). Each boxplot is the result of $B = 1000$ simulations. Properties of these distributions are summarized in Table 1.

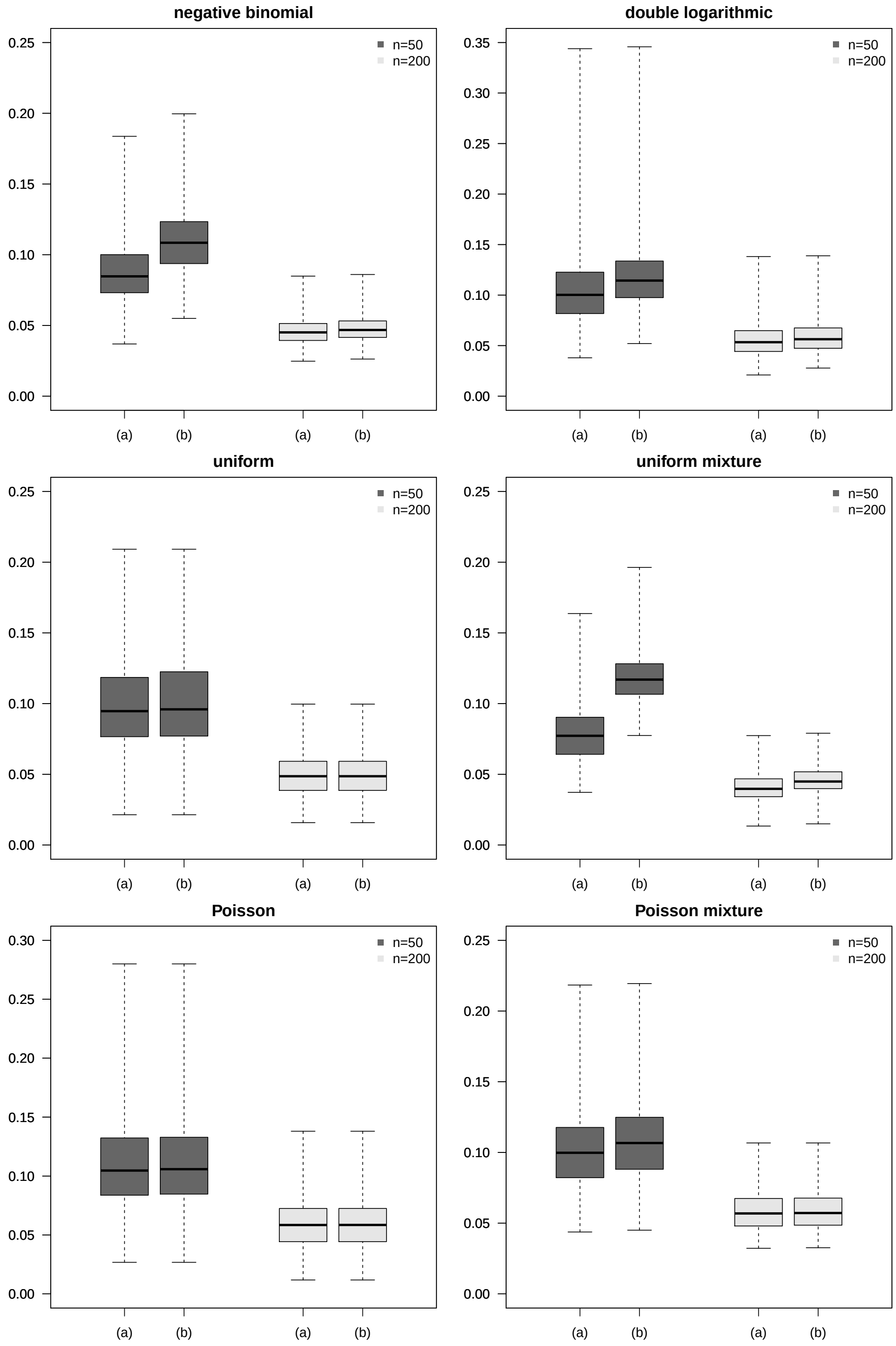

Figure 3: Boxplots for the six distributions of Figure 2 of the ℓ_2 distance of the estimated pmf from the true pmf for the unimodal MLE when the support is known (a) and unknown (b). Each boxplot is the result of $B = 1000$ simulations.

Properties of the six distributions are summarized in Table 1 for convenience.

In Figure 2, we can see that the unimodal MLE performs better than the empirical MLE for all six distributions. It is, however, outperformed by the log-concave MLE for the distributions which are log-concave, although they seem to have very comparable errors in the case of the Poisson distribution with rate $\lambda = 2$. On the other hand, the unimodal MLE outperforms the log-concave MLE when the distribution is not log-concave, at least for sample sizes which are "large enough". Our simulations show that this sample size is related to how far away the true pmf is from the set of log-concave distributions. In Figure 2, the ℓ_2 distance to the corresponding log-concave Kullback-Leibler projection (cf. Balabdaoui et al. (2013)) is approximatively 0.363 for the double logarithmic and 0.050 for the uniform mixture. Overall, we expect that when log-concavity fails to hold, the unimodal MLE will be the better estimator for larger sample sizes. Moreover, this behavior will hold also for smaller sample sizes for pmfs that are further away from the log-concave class. The bimodal Poisson mixture model is the only example in which neither the log-concave nor unimodal classes are correct. Notably, although the empirical pmf is the only well-specified MLE in this case, it outperforms the other two estimators only for the largest sample size.

3.1 Comparison of known versus unknown support

It seems self-evident that some efficiency will be lost by assuming that the support is unknown. Here, we briefly consider the question of "how much efficiency is lost?" via simulations. To be precise, when we say that the support is known, the MLE is defined as

$$\text{argmax}_{p \in \mathcal{U}^1(S)} \ L_n(p),$$

unlike in the definition of $\widehat{p}_n$, where S is replaced by its estimate S_n. In order to avoid existence issues of the estimator defined above, the class $\mathcal{U}^1(S)$ should be viewed as the set of probability mass functions p with support contained in S. Our simulations show that although some difference is seen for small a sample size, the cost is not great, and the difference disappears with increased sample size. As mentioned previously, our consistency and asymptotic results developed later apply to both versions of the MLE.

Figure 4 gives an example of the two approaches (known vs. unknown support in the unimodal MLE) for a sample from the negative binomial distribution for $n = 50$. Both unimodal MLE approaches provide considerable "smoothing" to the empirical pmf. However, when the support is unknown, the MLE will only place mass on S_n, whereas the MLE with known support will place mass on the entire range $\{X_{(1)}, \ldots, X_{(n)}\}$. This is clearly seen in Figure 4. Thus, the potential loss of efficiency will most likely occur in the tails of the true distribution, and this will be particularly true for distributions with a fatter tail.

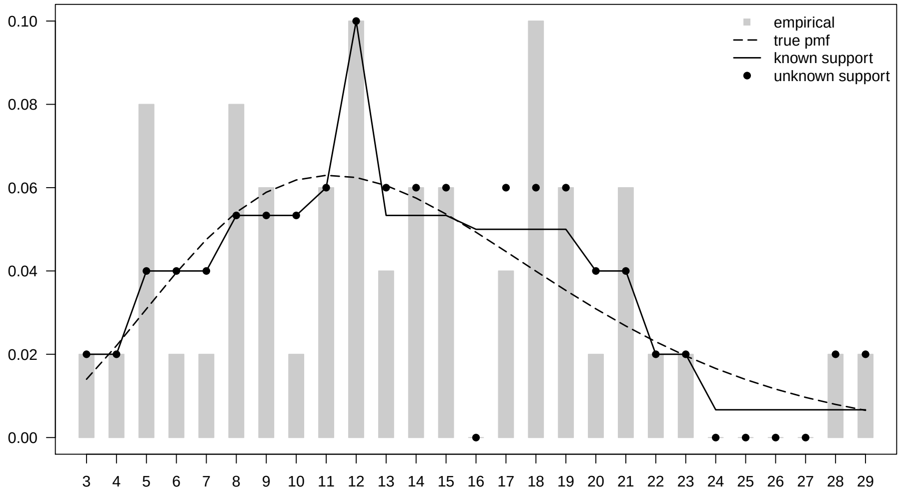

Figure 4: Example comparing the unimodal MLE when the support is known vs. unknown. The true distribution is the negative binomial with sample size $n = 50$.

We compared the two approaches via simulations, the results of which are shown in Figure 3. The distributions considered are exactly the same as those described on page 8. The loss is small for the uniform distribution which has support on only ten points, and also for both Poisson distributions, where the tails converge to zero very quickly. For the other distributions, which slower rate of decay in the tails, some efficiency is lost for the small sample size ($n = 50$). However, the loss appears almost negligible for the medium sample size ($n = 200$).

Remark 3.1. *When $S = S_n$, the known support and unknown support MLE versions will be the same. For the case when $|S|$ is finite, the probability that this does not happen for a given n decreases exponentially with n. Furthermore, with probability one, there exists an n_0, such that for all $n \geq n_0$, $S = S_n$ in this case.*

4 The Kullback-Leibler projection and consistency of the unimodal MLE

Let p_0 denote a fixed probability mass function on S_0 with distribution function P_0. We let

$$\rho(p|p_0) \quad = \quad \int \log \frac{p_0}{p} \, dP_0,$$

denote the Kullback-Leibler (KL) divergence. In this section, we seek the KL projection $\widehat{p}_0 \in \mathcal{U}^1(S_0)$ of a given pmf p_0. The KL projection has been considered extensively for the log-concave shape constraint for densities on $\mathbb{R}^d$ in Cule and Samworth (2010) and Dümbgen et al. (2011) and for probability mass functions in Balabdaoui et al. (2013). As in Cule and Samworth (2010); Cule et al. (2010); Balabdaoui et al. (2013), we can define such a projection as

$$\widehat{p}_0 \quad = \quad \mathrm{argmin}_{p \in \mathcal{U}^1(S_0)} \int_{S_0} \log \frac{p_0}{p} \, dP_0 \quad = \quad \mathrm{argmin}_{p \in \mathcal{U}^1(S_0)} \rho(p|p_0), \tag{4.8}$$

which is the element of $\mathcal{U}^1(S_0)$ closest to the unknown pmf p_0 in the sense of Kullback-Leibler divergence. From a practical point of view, this allows us to view the shape constrained estimator as the closest approximation within a class of distributions.

Alternatively, Patilea (2001) uses the definition

$$\int \log \frac{\widehat{p}_0}{p} dP_0 \quad \geq \quad 0, \qquad \text{for all } p \in \mathcal{U}^1(S_0), \tag{4.9}$$

and refers to the pmf $\widehat{p}_0$ satisfying (4.9) as the pseudo-true pmf. If the integrals involved are finite, one can re-arrange (4.9) into (4.8) and vice versa. In particular, if $\inf_{q \in \mathcal{U}^1(S_0)} \rho(q|p_0) = \rho(\widehat{p}_0|p_0) < \infty$ for some $\widehat{p}_0$ then (4.8) is equivalent to (4.9), since then

$$\begin{aligned} 0 \leq \int \log \frac{p_0}{\widehat{p}_0} dP_0 \quad &\leq \quad \int \log \frac{p_0}{p} dP_0 \quad = \quad \int \log \frac{p_0}{\widehat{p}_0} \frac{\widehat{p}_0}{p} dP_0 \\ &= \quad \int \log \frac{p_0}{\widehat{p}_0} dP_0 + \int \log \frac{\widehat{p}_0}{p} dP_0. \end{aligned}$$

Alternatively, as in Dümbgen et al. (2011), one could also consider

$$\int \log \widehat{p}_0 \, dP_0 \quad \geq \quad \int \log p \, dP_0, \qquad \text{for all } p \in \mathcal{U}^1(S_0), \tag{4.10}$$

which is akin to maximizing the likelihood. If p_0 admits a finite entropy, that is $\int \log p_0 \, dP_0 > -\infty$, then (4.10) is equivalent to (4.8). Furthermore, (4.9) is equivalent to (4.10) whenever $\sup_{p \in \mathcal{U}^1(S_0)} \int \log p \, dP_0 > -\infty$ and is attained.

In what follows, we work with the formulation of Patilea (2001) in (4.9), although we continue to refer to it as the KL projection. Before stating our first theorem, we recall that $\mathcal{U}^1|_\kappa(S_0)$ is the space of unimodal pmfs with support S_0 and mode at either $s_{\kappa-1}$ or s_κ.

Theorem 4.1. *Let p_0 be a discrete pmf with support S_0. Let $\widehat{P}_0|_\kappa$ denote the greatest convex majorant of the cumulative sum of $p_0(s_i), i \leq \kappa - 1$ and the least concave minorant of the cumulative sum of $p_0(s_i), i \geq \kappa$, and let $\widehat{p}_0|_\kappa$ denote the pmf corresponding to $\widehat{P}_0|_\kappa$. Then*

$$\int \log \frac{\widehat{p}_0|_\kappa}{p} dP_0 \quad \geq \quad 0, \qquad \textit{for all } p \in \mathcal{U}^1|_\kappa(S_0). \tag{4.11}$$

Furthermore, when $p_0 \in \mathcal{U}^1|_\kappa(S_0)$*, or when* $\sum_{j\neq 0} \log|j| p_0(s_j) < \infty$*,* $q = \widehat{p_0}|_\kappa$ *is the unique pmf which satisfies* $\int \log(q/p) dP_0 \geq 0$ *for all* $p \in \mathcal{U}^1|_\kappa(S_0)$*.*

We next consider the larger class $\mathcal{U}^1(S_0)$.

Theorem 4.2. *Let* p_0 *be a discrete pmf with support* S_0*.*
1. Suppose that $p_0 \in \mathcal{U}^1(S_0)$*. Then* $\widehat{p_0} = p_0$ *is the unique unimodal pmf satisfying*

$$\int \log \frac{\widehat{p_0}}{p} dP_0 \quad \geq \quad 0 \qquad \textit{for all } p \in \mathcal{U}^1(S_0).$$

2. Suppose that $p_0 \notin \mathcal{U}^1(S_0)$ *and* $\sum_{j\neq 0} \log|j| p_0(s_j) < \infty$*. Then there exists a* $\widehat{p_0} \in \mathcal{U}^1(S_0)$ *such that*

$$\int \log \frac{\widehat{p_0}}{p} dP_0 \quad \geq \quad 0 \qquad \textit{for all } p \in \mathcal{U}^1(S_0).$$

When $\widehat{p_0}$ *is not unique, we shall denote by* $\{\widehat{p_0}\}$ *the (finite) collection of all such projections.*

Theorem 4.2 says that in case the model is well-specified, then the KL projection of p_0 is unique and equal to the true pmf itself under no additional assumptions. However, if the model is misspecified, there may exist several different KL projections. These are collected in the set $\{\widehat{p_0}\}$ which is necessarily finite. Examples of such non-uniqueness are given later in Figure 5. We believe that this lack of uniqueness is due to the fact that the space of unimodal densities is not convex. Although the condition $\sum_{j\neq 0} \log|j| p_0(s_j) < \infty$ may seem a bit unnatural at first, one can express it in a more transparent form thanks to the next proposition.

Proposition 4.3. *Let* p_0 *be a discrete pmf with support* S_0*. Then*

$$\sum_{j\neq 0} \log|j| \; p_0(s_j) < \infty \quad \textit{if and only if} \quad \sup_{p\in\mathcal{U}^1(S_0)} \int \log p \; dP_0 \in (-\infty, 0].$$

Recall that under this condition, (4.9) is equivalent to (4.10). Furthermore, if we assume in addition that

$$0 < \delta_1 \quad \leq \quad \inf(s_{j+1} - s_j) \quad \leq \quad \sup(s_{j+1} - s_j) \quad \leq \delta_2 < \infty,$$

then one can show that the condition $\sum_{j\neq 0} \log|j| p_0(s_j) < \infty$ is equivalent to $\int \log|x - a| dP_0(x) \in \mathbb{R}$ for some $a \notin S_0$. Therefore, this condition gives a bound on the speed of decay of p_0. Also, it is weaker than the assumption of having a finite mean required by Cule and Samworth (2010); Dümbgen et al. (2011). Our assumption is also weaker than that made by Patilea (2001, Corollary 5.6), although the latter is a condition in order to derive rates of convergence. In our setting, Patilea's assumption boils down to existence of an $\epsilon > 0$ such that

$$\int \widehat{p_0}^{-\epsilon} dP_0 < \infty$$

where P_0 is the cumulative distribution function of p_0. By the inequality $\log(x) \leq x^\epsilon/\epsilon$ for $x \in (0, \infty)$, we find

$$\int \log(1/\widehat{p_0})\, dP_0 \leq \frac{1}{\epsilon} \int \widehat{p_0}^{-\epsilon} dP_0 < \infty,$$

implying our condition in Proposition 4.3 (since then $\int \log \widehat{p_0} dP_0 > -\infty$).

4.1 Consistency

For two pmfs p and q defined on S, let $\ell_k(p,q)$ and $h(p,q)$ denote the ℓ_k and Hellinger distances between p and q, respectively. That is,

$$\ell_k(p,q) = \begin{cases} \left(\sum_{x\in\mathcal{S}} |p(x) - q(x)|^k\right)^{1/k}, & \text{if } 1 \leq k < \infty \\ \sup_{x\in\mathcal{S}} |p(x) - q(x)|, & \text{if } k = \infty \end{cases}$$

and

$$h(p,q) = \frac{1}{2} \sum_{k\in\mathcal{S}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2.$$

In the following, we establish almost sure consistency of the unimodal MLE under a mild condition on the true pmf p_0. Let us fix a discrete pmf p_0 with support S_0, and assume that we observe i.i.d. data $X_1, \ldots, X_n \sim p_0$. Here, we do not necessarily assume that p_0 is itself unimodal. Let $\widehat{p}_n$ denote again the unimodal MLE based on the sample $(X_1, \ldots, X_n)$. Recall that in the well-specified model, the KL projection $\widehat{p_0}$ in the sense of (4.9) is p_0 itself. When the model is misspecified and p_0 satisfies $\sum_j \log|j| p_0(s_j) < \infty$, then the KL projection $\widehat{p_0}$ exists in the sense of (4.10) but may not be unique. In this situation, we denote by $\{\widehat{p_0}\}$ the set of all such KL projections.

Theorem 4.4. *Suppose that $\sum_{i\neq 0} \log|i| p_0(s_i) < \infty$, and let $d \equiv \ell_k$ or h. Then*

$$d(\widehat{p}_n, \{\widehat{p_0}\}) \equiv \inf_{\widehat{q}\in\{\widehat{p_0}\}} d(\widehat{p}_n, \widehat{q}) \to 0$$

almost surely. If p_0 is unimodal, then

$$d(\widehat{p}_n, p_0) \to 0$$

almost surely.

Remark 4.5. *Pointwise convergence and convergence in $\ell_k, 1 \leq k \leq \infty$ and Hellinger distance h are all equivalent for probability mass functions. This follows for example from Lemma C.2 in the online supporting material of Balabdaoui et al. (2013).*

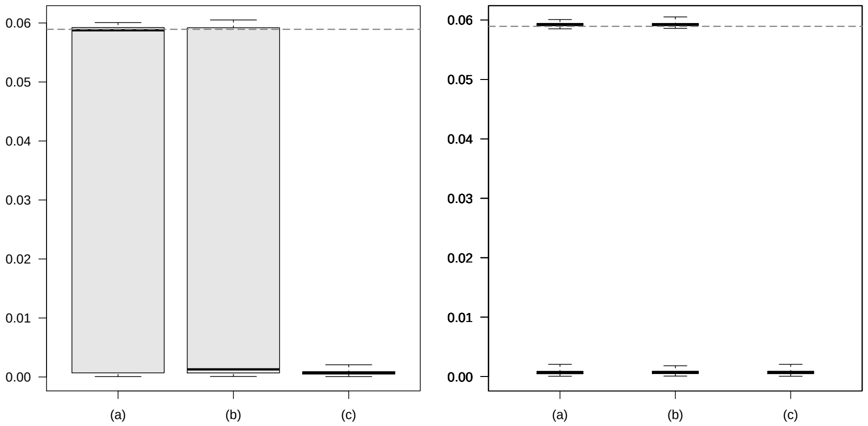

Figure 5: Convergence of $\widehat{p}_n$ to $\{\widehat{p}_0\}$ for p_0 as in (4.12). The boxplots show the $d = \ell_2$ distance for $B = 1000$ Monte Carlo samples with a sample size of $n = 1\,000\,000$. The three columns give (a) $d(\widehat{p}_n, \widehat{p}_0^1)$, (b) $d(\widehat{p}_n, \widehat{p}_0^2)$, and (c) $d(\widehat{p}_n, \{\widehat{p}_0\})$. The plot on the right differs from the plot on the left in that, on the right, in (a) and (b) the boxplots have been split into large/small values to show the bimodal nature of the data. For reference, the dashed horizontal line gives $d(\widehat{p}_0^1, \widehat{p}_0^2)$.

The fact that $\{\widehat{p}_0\}$ is not necessarily a singleton means that the MLE does not necessarily converge to a particular element of $\{\widehat{p}_0\}$. Rather, our proof shows instead that the MLE is sequentially compact: there exists an element $\widehat{q} \in \{\widehat{p}_0\}$ and a subsequence n_k such that $d(\widehat{p}_{n_k}, \widehat{q}) \to 0$. We illustrate this behaviour via the following example. Let $S_0 = \{-2, -1, 0, 1, 2\}$ and define

$$p_0(s_i) \quad = \quad \begin{cases} 1/6 & s_i = -2, 0, 2 \\ 1/4 & s_i = -1, 1. \end{cases} \tag{4.12}$$

In this case, $\{\widehat{p}_0\}$ has two elements, which we denote by $\widehat{p}_0^1$ and $\widehat{p}_0^2$. Straightforward calculations show that

$$\widehat{p}_0^1(s_i) = \begin{cases} 1/6 & s_i = -2, 2 \\ 1/4 & s_i = -1, \\ 5/24 & s_i = 0, 1 \end{cases}$$

with mode at -1 and $\widehat{p}_0^2(s_i) = \widehat{p}_0^1(-s_i)$ (with mode at 1). Simulations for a very large sample size are shown in Figure 5, where the convergence in set distance is clearly visible.

On the other hand, if $|\{\widehat{p}_0\}| = 1$, then the unimodal MLE converges to the unique element of $\{\widehat{p}_0\}$. We also note that if we consider the restricted MLE $\widehat{p}_n|_\kappa$, then a similar result to the above holds. A proof may be provided using, for example, Marshall's lemma as in Patilea (2001, Lemma 5.5, page 114), without any restrictions on p_0.

Let $\widehat{\mathcal{M}}_n$ be the modal region of the unimodal MLE $\widehat{p}_n$, cf. (2.2). An immediate corollary of the preceding theorem is the following statement about convergence of $\widehat{\mathcal{M}}_n$. For simplicity, we now assume that the KL projection is unique. However, one may state the following results with some additional generality, albeit in a less clear manner.

Corollary 4.6. *Assume $\sum_{i\neq 0} \log|i| p_0(s_i) < \infty$ and that $|\{\widehat{p}_0\}| = 1$, and let $\mathcal{M}$ denote the modal region of p_0. Then with probability one, there exists a sufficiently large n_0, such that for all $n \geq n_0$, $\widehat{\mathcal{M}}_n \subset \mathcal{M}$.*

If $|\mathcal{M}| = 1$, then Corollary 4.6 implies that, with probability one, there exists a sufficiently large n_0, such that the mode of the MLE coincides with the true mode. In the case $|\mathcal{M}| > 1$, then this is no longer true, and all we can say is that eventually the estimated mode will be in $\mathcal{M}$. Note that the latter has nothing to do with the fact that we make the convention of taking the smallest mode to define the MLE; any such convention (or no convention) would result in the same behavior.

From Theorem 4.4 we immediately obtain the following result.

Corollary 4.7. *Assume $\sum_{i\neq 0} \log|i| p_0(s_i) < \infty$ and that $|\{\widehat{p}_0\}| = 1$. Let $\widehat{F}_n$ and $\widehat{F}_0$ denote the cdfs of $\widehat{p}_n$ and $\widehat{p}_0$ respectively. Then*

$$\lim_{n\to\infty} \sup_{s\in S_0} |\widehat{F}_n(s) - \widehat{F}_0(s)| = 0$$

almost surely.

5 Global asymptotics

The asymptotic behaviour of the unimodal MLE, as well as the proof thereof, share many similarities with those given in Jankowski and Wellner (2009) for the Grenander estimator of decreasing pmf on $\mathbb{N}$. Our main interest here is to derive the weak limit of the estimator when p_0 is unimodal, and therefore we do not consider the misspecified setting. One could, however, mimic the work in Jankowski (2014) to obtain the asymptotic distributions in this case under some further restrictions on p_0. Despite the similarity mentioned earlier with the monotone problem, some technical details need special attention due to the fact that (1) the mode of the true pmf is unknown, and (2) we do not assume that the true support is known.

To describe the asymptotic theory, we first need to define an operator, denoted here as φ. Recall that $\mathcal{M}$ denotes the modal region of p_0 as defined in (2.2). Let us write

$$\begin{aligned} D &= \Big\{ s_i : s_i \notin \mathcal{M} \text{ and } p_0(s_i) \geq p_0(s_{i+1}) \Big\}, \quad \text{and} \\ I &= \Big\{ s_i : s_i \notin \mathcal{M} \text{ and } p_0(s_{i-1}) \leq p_0(s_i) \Big\} \end{aligned}$$

as the decreasing and increasing regions of S_0 respectively. We will write $\mathcal{M} = \{\tau_0^I, \dots, \tau_0^D\}$ (where $\tau_0^I \le \tau_0^D$), and let $\{\tau_i^D\}_{i\ge 1}$ enumerate the points in D such that $p_0(s_i) > p_0(s_{i+1})$, where $\tau_i^D < \tau_{i+1}^D$. Similarly, let $\{\tau_i^I\}_{i\ge 1}$ enumerate the points in I such that $p_0(s_{i-1}) < p_0(s_i)$, where $\tau_{i+1}^I < \tau_i^I$. We will write $D_j = \{s \in S_0, \tau_{j-1}^D < s \le \tau_j^D\}$ for $j \ge 1$, and $I_j = \{s \in S_0, \tau_j^I \le s < \tau_{j-1}^I\}$ for $j \ge 1$. Notice that each of these regions is necessarily finite, and that p_0 is constant on each subset I_j, D_j and $\mathcal{M}$. We therefore have that

$$I = \cup I_j, \; D = \cup D_j, \text{ and } \mathcal{S}_0 = I \cup \mathcal{M} \cup D. \tag{5.13}$$

We also denote the collection of knots as

$$\mathcal{T} = \{\tau_j^I, j \ge 1\} \cup \{\tau_0^D, \tau_0^I\} \cup \{\tau_j^D, j \ge 1\}. \tag{5.14}$$

Note that our definition of a knot, as well as the collection of knots, depends on the underlying pmf p_0. Finally, let q be an element of $\ell_2(S_0)$, and for a subset $C \subset S_0$ we write the vector $q_C = \{q(s_j), s_j \in C\}$ to denote the sequence q restricted to C. We may now define φ :

$$\varphi[q](s) = \begin{cases} \mathrm{iso}[q_{I_j}](s) & s \in I_j, \\ \mathrm{uni}[q_{\mathcal{M}}](s) & s \in \mathcal{M}, \\ \mathrm{anti}[q_{D_j}](s) & s \in D_j. \end{cases} \tag{5.15}$$

Note that the definition of φ technically depends on p_0, although we omit this dependence in the notation. In addition, φ satisfies $\varphi[p_0] = p_0$.

Theorem 5.1. *Suppose that p_0 is unimodal and that $\sum_{i\neq 0} \log|i| p_0(s_i) < \infty$. Let $\mathbb{W}$ denote the discrete white noise process: That is, $\mathbb{W}$ is the mean zero Gaussian process defined on S_0 such that cov$(\mathbb{W}(s_i), \mathbb{W}(s_j)) = p_0(s_i)\delta_{i,j} - p_0(s_i)p_0(s_j)$. Then*

$$\sqrt{n}(\widehat{p}_n - p_0) \Rightarrow \varphi[\mathbb{W}],$$

in $\ell_k(S_0)$, where $2 \le k \le \infty$.

An immediate corollary of our result is that if s is such that $s \in C$ where $C = I_j, \mathcal{M}$, or D_j and $|C| = 1$, then $\sqrt{n}(\widehat{p}_n(s) - p_0(s)) \Rightarrow \mathbb{W}(s)$, since in such cases $\varphi[q](s) = q(s)$. Namely, this says that in regions where p_0 is strictly unimodal, the asymptotics of $\widehat{p}_n$ are the same as those of $\overline{p}_n$. Similar observations have been made in Jankowski and Wellner (2009) for the Grenander estimator and Balabdaoui et al. (2013) for the log-concave MLE. In addition, we note that $\ell_2(S_0)$ is the smallest space, of those considered above, where one can prove the asymptotics. In other words, convergence in a smaller space such as $\ell_1(S_0)$ cannot be considered without additional assumptions on p_0. We refer to Jankowski and Wellner (2009) for additional details.

The next result follow immediately from the definition of the operator φ as well as Jankowski and Wellner (2009, Theorem 2.1).

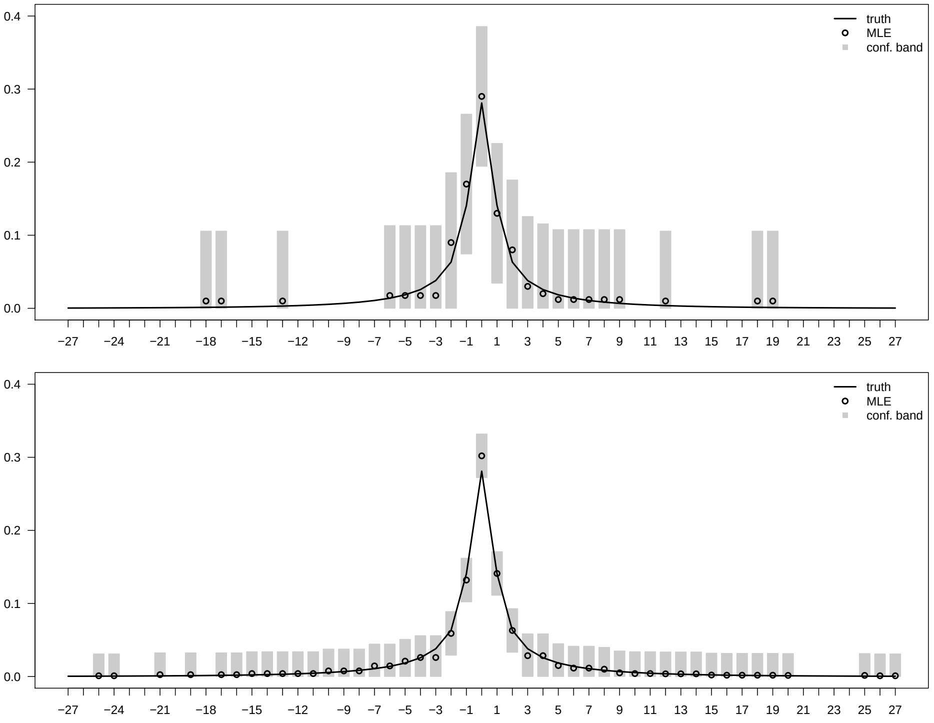

Figure 6: 95% constant-width ($\beta = 0$) confidence bands for the true pmf when sampling from the double logarithmic distribution $p = 0.9$. The sample size is $n = 100$ on the top and $n = 1000$ on the bottom.

Proposition 5.2. *For $2 \le k \le \infty$, we have that $\|\varphi[\mathbb{W}]\|_k \le \|\mathbb{W}\|_k$.*

In addition, it is also possible to develop a Marshall's lemma type result in our setting. The (asymptotically negligible) error term not seen in the usual type of result here is due to estimation of the support in our approach.

Proposition 5.3 (Marshall's Lemma)**.** *Suppose that $\sum_{s \in S_0} p_0^{1/2}(s) < \infty$ and that the true pmf p_0 is unimodal with associated cumulative distribution function F_0. Then, with probability one, there exists an n_0 such that for all $n \ge n_0$*

$$\sup_{s \in S_0} |\widehat{F}_n(s) - F_0(s)| \le \sup_{s \in S_0} |\mathbb{F}_n(s) - F_0(s)| + o_p(n^{-1/2}).$$

6 Global confidence bands for p_0

The key application of the previous section is the calculation of confidence bands for the true pmf p_0, which we assume to be unimodal. To this end, let $q_{0,\alpha}$ be such that $P(\|\mathbb{W}\|_\infty > q_{0,\alpha}) = \alpha$. Then, it follows that

$$\lim_n P\left(\sqrt{n}\|\widehat{p}_n - p_0\|_\infty \le q_{0,\alpha}\right) \quad \ge \quad 1 - \alpha.$$

This follows since

$$\sqrt{n}\|\widehat{p}_n - p_0\|_\infty \quad \Rightarrow \quad \|\varphi[\mathbb{W}]\|_\infty \quad \le \quad \|\mathbb{W}\|_\infty.$$

It is important to note that if p_0 is strictly monotone then $\varphi[\mathbb{W}] = \mathbb{W}$, and then the last inequality above becomes an equality, resulting in an asymptotically *exact* confidence band.

In order to estimate $q_{0,\alpha}$, we use $\widehat{p}_n$ in place of p_0. In Proposition B.7, we show that this yields an almost surely consistent method of estimating $q_{0,\alpha}$. Also, we estimate each quantile using Monte Carlo simulations. Thus, let $\widehat{q}_{0,\alpha}$ denote the Monte Carlo estimate of the quantile of $\|\mathbb{W}\|_\infty$.

It follows that an asymptotically correct conservative confidence band is given by

$$\left\{\left[\left(\widehat{p}_n(s_i) - \frac{\widehat{q}_{0,\alpha}}{\sqrt{n}}\right) \vee 0,\ \widehat{p}_n(s_i) + \frac{\widehat{q}_{0,\alpha}}{\sqrt{n}}\right], s_i \in \operatorname{supp}(\widehat{p}_n)\right\} \tag{6.16}$$

where $\operatorname{supp}(\widehat{p}_n)$ denotes the support of $\widehat{p}_n$. When the support of p_0 is estimated from the data, then $\operatorname{supp}(\widehat{p}_n) = S_n$, the support of the empirical distribution.

In Figure 6, we show an example of confidence bands thus constructed, when the true pmf is the double logarithmic distribution with $p = 0.9$. We found the constant width of the confidence bands, particularly for the smaller sample size, somewhat visually jarring. For this reason, we also create confidence bands which are visually more appealing in that they do not have uniform width. Define, for $\beta \ge 0$,

$$\widehat{\mathbb{W}}_n^\beta(s) = \begin{cases} \frac{\sqrt{n}(\widehat{p}_n - p_0)(s)}{\widehat{p}_n^\beta(s)}, & s \in \operatorname{supp}(\widehat{p}_n) \\ 0, & s \notin \operatorname{supp}(\widehat{p}_n). \end{cases}$$

If $\beta = 0$, then $\widehat{\mathbb{W}}_n^\beta = \sqrt{n}(\widehat{p}_n - p_0)$, and we are in the situation of constant-width confidence bands discussed above.

Proposition 6.1. *Fix $\beta > 0$ and assume that the support of p_0 is finite. Then*

$$\|\widehat{\mathbb{W}}_n^\beta\|_\infty \quad \Rightarrow \quad \left\|\frac{\varphi[\mathbb{W}]}{p_0^\beta}\right\|_\infty \quad \le \quad \left\|\frac{\mathbb{W}}{p_0^\beta}\right\|_\infty.$$

In this case, an asymptotically correct conservative confidence band is given by

$$\left\{\left[\left(\widehat{p}_n(s) - \widehat{p}_n^\beta(s)\frac{\widehat{q}_{\beta,\alpha}}{\sqrt{n}}\right) \vee 0,\ \widehat{p}_n(s) + \widehat{p}_n^\beta(s)\frac{\widehat{q}_{\beta,\alpha}}{\sqrt{n}}\right], s \in \operatorname{supp}(\widehat{p}_n)\right\},$$

where $\widehat{q}_{\beta,\alpha}$ is an estimate of $q_{\beta,\alpha}$ where $P\left(\left\|p_0^{-\beta}\mathbb{W}\right\|_\infty > q_{\beta,\alpha}\right) = \alpha$. Estimation of this quantile can be done using a Monte Carlo approach, as before.

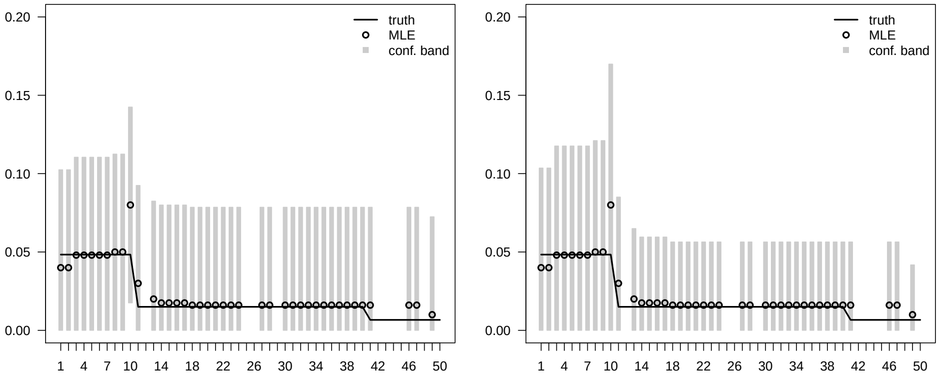

Figure 7: 95% confidence bands for the true pmf when sampling from the mixture of uniforms distribution with mixing distribution given in (3.7). The sample size is $n = 100$ and we chose $\beta = 0$ (left, constant width) and $\beta = 0.5$ (right, varying width).

Remark 6.2. *When p_0 has infinite support, the limiting distribution $p_0^{-\beta}\mathbb{W}$ exists in ℓ_2 provided that $\sum p_0^{1-2\beta} < \infty$, which adds the restriction that $\beta < 1/2$. We conjecture that the above result continues to hold for distributions with infinite support with the restriction that $\beta \in [0, 1/2)$, although we do not pursue the proof here. We do note that the assumption of finite support may be highly plausible in certain practical situations, whereas the (weaker) assumption of $\sum_i p_0^{1-2\beta} < \infty$, may not be as easy to motivate.*

In Figure 7 we compare the constant-width confidence band to the varying width confidence band (with $\beta = 0.5$) when the true distribution is the mixture of uniforms, whose mixing distribution is given in (3.7). Visually, we find the choice of $\beta = 0.5$ preferable in that the values, where $\widehat{p}_n$ is smaller, express slightly more accuracy, as one would expect. In this particular example, the difference is not great, but is still eye-pleasing. For $\beta = 0$, the width of the confidence bands varies from 0.13 to 0.07 (median 0.08), while for $\beta = 0.5$, the width of the confidence bands varies from 0.17 to 0.04 (median 0.06). Note that, although for $\beta = 0$ the confidence bands have constant width, we have to cut off the lower bound at a maximum value of zero, and hence the bands end up being non-constant in reality. Without this cutoff, the width would be constant at 0.13.

In Table 2, we examine the empirical performance of the proposed confidence bands. We consider two different unimodal distributions: the mixture of uniforms as above, and the double logarithmic with $p = 0.9$ from (3.6). Our simulations span various samples sizes and values of β. Note that when $\beta = 0.5$, and the true pmf is double logarithmic, the conditions for convergence are violated (see Proposition 6.1 and Remark 6.2), and we include this example for comparison only (seeing as the condition that $\sum_i p_0^{1-2\beta} < \infty$ may be difficult to verify without additional information about p_0).

Table 2: Empirical coverage probabilities for the proposed confidence bands with $\alpha = 0.05$.

	β	$n = 100$	$n = 1000$	$n = 5000$
mixture of uniforms	0	0.972	0.963	0.959
	0.25	0.991	0.971	0.970
	0.5	0.959	0.953	0.991
double logarithmic	0	0.956	0.949	0.949
	0.25	0.970	0.950	0.948
	0.5	0.980	0.989	0.989

Define, for $\beta \geq 0$,

$$\begin{aligned} \widehat{c}_{n,u}(s) &= \widehat{p}_n(s) + \widehat{p}_n^{\beta}(s)\frac{\widehat{q}_{\beta,\alpha}}{\sqrt{n}}, \ s \in \mathrm{supp}(\widehat{p}_n), \\ \widehat{c}_{n,l}(s) &= 0 \vee \left(\widehat{p}_n(s) - \widehat{p}_n^{\beta}(s)\frac{\widehat{q}_{\beta,\alpha}}{\sqrt{n}}\right), \ s \in \mathrm{supp}(\widehat{p}_n), \\ \widehat{c}_{n,u}(s) &= \widehat{c}_{n,l}(s) \ = \ 0, \ s \notin \mathrm{supp}(\widehat{p}_n). \end{aligned}$$

The results in Table 2 give the empirical coverage *on the set* S_n as indicated in the third column. That is, we report the proportion of times that

$$\widehat{c}_{n,l}(s) \ \leq \ p_0(s) \ \leq \ \widehat{c}_{n,l}(s), \quad \text{for all } s \in S_n \tag{6.17}$$

was observed.

Overall, we find that the confidence bands perform rather well. Note that the double logarithmic case $\beta < 0.5$ we would expect to obtain asymptotically correct bands, whereas in both uniform mixture scenarios, we expect an asymptotically conservative result. In Appendix A, we provide some additional results where we study the cost of defining the bands on $\mathrm{supp}(\widehat{p}_n) = S_n$ in the simulations.

7 Time-to-onset of the Ebola virus

In a recent article, Breman and Johnson (2014) describe their experiences during the 1976 Ebola virus outbreak in Zaire (currently, the Democratic Republic of the Congo). The figure in the article shows histograms of the time of onset of the disease based on the transmission route: patients became infected either with an unsterilized needle or through person-to-person contact. This data was also previously published in Breman et al. (1978). Here, we use the histograms in Breman and Johnson (2014) to transcribe the

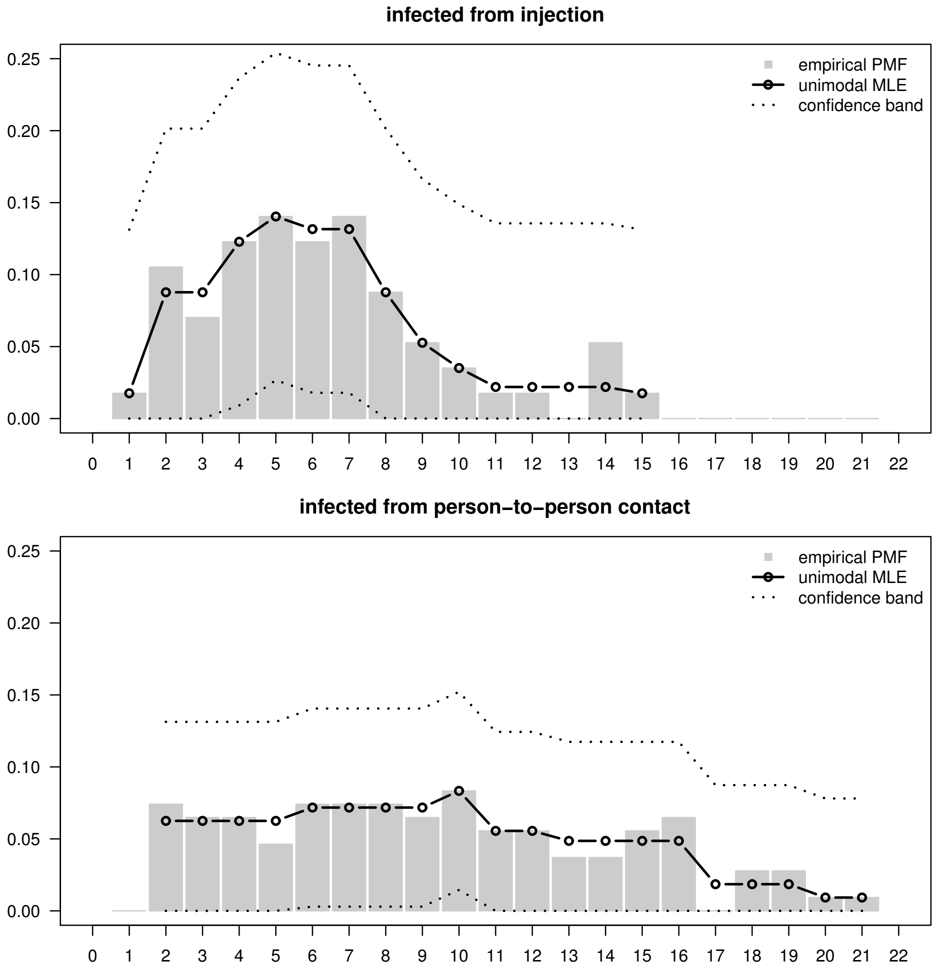

Figure 8: Time to onset of symptoms of the Ebola virus based on transmission type. The sample size is $n = 57$ for those infected from unsterilized needles and $n = 108$ for person to person contact.

data and perform a brief analysis. We note that transcribing the histograms resulted in samples sizes of $n = 57$ and $n = 108$, which differs slightly from that presented in Breman and Johnson (2014).

Figure 8 shows the empirical observations and the fitted unimodal MLEs for the two types of transmission routes. The 95% asymptotic global confidence band are also included, where we used the version with constant width; i.e. those corresponding to $\beta = 0$. Note that the time-to-onset is measured in days, and it therefore makes sense to assume that the support of the true pmf is either equal to the natural numbers, or is a connected subset thereof. Thus, we use the version of the likelihood maximization where the support is not estimated from the data. As mentioned earlier, our results apply also to this (easier) case. Visually, there is no glaring reason that the assumption of unimodality is not appropriate in these two cases. On the other hand, the fitted MLE provides a slight smoothing to the empirical distribution, which is appealing.

We note also that the confidence bands in Figure 8 appear somewhat wide at first glance. However, this is due to the smaller sample sizes observed in both distributions. The average width for the injection infection was found to be 0.18, and 0.12 for infection from person-to-person contact. As a crude benchmark, the average widths of 95% *pointwise* confidence intervals were calculated for the true pmf

$$\widehat{p}_n \pm 1.96\,\sqrt{\widehat{p}_n(1-\widehat{p}_n)},$$

based on Theorem 5.1 and under the (untested) assumption that the true pmf is strictly unimodal. Here, the average width for the injection infection was found to be 0.12, and 0.08 for infection from person-to-person contact. These are also rather wide, but smaller than the global confidence bands, as expected.

It is quite interesting how different the two distributions appear to be. The standard Kolmogorov-Smirnov test does not yield exact p-values in this setting because the data is discretized, and hence we used a permutation test (Jöckel, 1986). This modified approach yielded a p-value of 0.0014 for the hypothesis that the two distributions are the same (incorrectly applying the regular Kolmogorov-Smirnov test also rejected the null hypothesis). This is in line with what we observe in Figures 8 and 9. R (R Core Team, 2014) code for performing this analysis is available online at `www.math.yorku.ca/~hkj/Research`.

A biological explanation for the difference between the two distributions was provided to us by Jane Heffernan (2014, private communication): "Injection gets the pathogen into the blood stream. Person-to-person contact provides exposure to the mucosa (innate immunity) first, so the pathogens that ultimately make it to the blood will be different in fitness distribution than the injection method. Also, the amount of pathogen ultimately making it to the blood could be smaller compared to the injection method. Both of these variables will affect the incubation period." In the data, we see this difference not only through a mean comparison (the mean time-to-onset is 6.3 days for transmission via injection and 9.4 days for person-to-person infections) but also in the stochastic dominance observed via the fitted and empirical CDFs in Figure 9. The latter suggest that $T_{inj} \leq T_{ptp}$ stochastically, where T_{inj} and T_{ptp} denote the times to onset for injection and person-to-person infections, respectively. Repeating the permutation test for the hypothesis that the two distributions are equal against the alternative that $F_{inj} > F_{ptp}$ yields a p-value of 0.0008.

Acknowledgments. The authors would like to thank Hélène Massam for

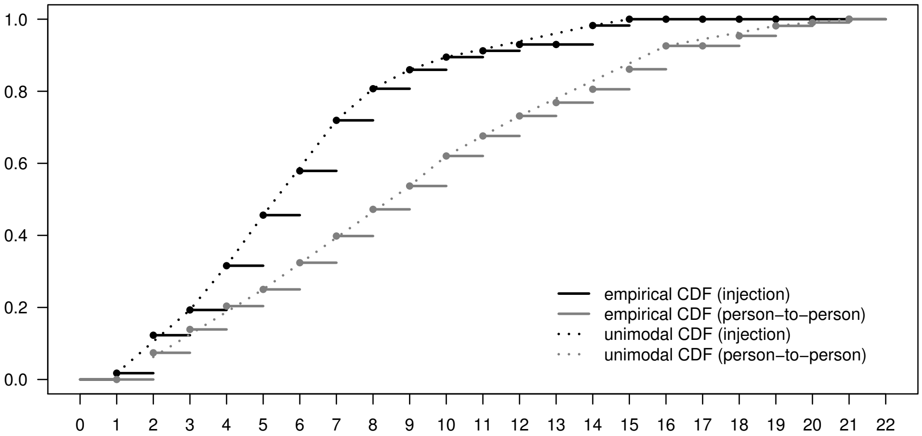

Figure 9: Time to onset of symptoms of the Ebola virus based on transmission type: a comparison of cumulative distribution functions

her help with the proof of Lemma B.5 and Jon Wellner for his help with the proof of Lemma B.6. We also thank the anonymous referees for useful comments which helped improve the paper.

References

Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymoptic distribution of the discrete log-concave mle and some applications. *JRSS, Series B* **75** 769–790.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25** 970–981.

Breman, J. G. and Johnson, K. M. (2014). Ebola then and now. *New England Journal of Medicine* **371** 1663–1666.

Breman, J. G., Piot, P., Johnson, K. M., White, M. W., Mbuyi, M., Sureau, P., Heymann, D. L., van Nieuwenhove, S., McCormick, J. B., Ruppol, J. P., Kintoki, V., Isaacson, M., van der Groen, G., Webb, P. A. and Ngvete, K. (1978). The epidemiology of ebola hemorrhagic fever in zaire, 1976. In *Pattyn SR, ed. Ebola virus hemorrhagic fever.* Elsevier, Amsterdam, 85–97.

Chowell, G., Bertozzi, S. M., Colchero, M. A., Lopez-Gatell, H., Alpuche-Aranda, C., Hernandez, M. and Miller, M. A. (2009). Severe respiratory disease concurrent with the circulation of H1N1 influenza. *New England Journal of Medicine* **361** 674–679.

Chowell, G., Fuentes, R., Olea, A., Aguilera, X., Nesse, H. and Hyman, J. (2013). The basic reproduction number R_0 and effectiveness of reactive interventions during dengue epidemics: The 2002 dengue outbreak in Easter Island, Chile. *Mathematical Biosciences and Engineering* **10** 1455–1474.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic J. Stat.* **4** 254–270.

Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 545–607.

Dudley, R. M. (1999). *Uniform central limit theorems*, vol. 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge.

Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function. *Bernoulli* **15** 40–68.

Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011). Approximation by log-concave distributions with applications to regression. *Ann. Statist.* **39** 702–730.

Durot, C., Huet, S., Koladjo, F. and Robin, S. (2013). Least-squares estimation of a convex discrete distribution. *Comput. Statist. Data Anal.* **67** 282–298.

Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153 (1957).

Harlan, S., Chowell, G., Yang, S., Petitti, D., Morales Butler, E., Ruddell, B. and Ruddell, D. M. (2014). Heat-related deaths in hot cities: Estimates of human tolerance to high temperature thresholds. *Int. J. Environ. Res. Public Health* **11** 3304–3326.

Jankowski, H. (2014). Convergence of linear functionals of the Grenander estimator under misspecification. *Annals of Statistics* To appear.

Jankowski, H. K. and Wellner, J. A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605.

Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Ann. Statist.* **14** 336–347.

Laskowski, M., Mostaco-Guidolin, L., Greer, A., Wu, J. and Moghadas, S. (2011). The impact of demographic variables on disease spread: influenza in remote communities. *Scientific Reports (Nature)* **1** 1–7.

Olshen, R. A. and Savage, L. J. (1970). A generalized unimodality. *J. Appl. Probability* **7** 21–34.

Patilea, V. (1997). *Convex models, NPMLE and misspecification.* Ph.D. thesis, Institute of Statistics, Univ. catholique de Louvain.

Patilea, V. (2001). Convex models, MLE and misspecification. *Ann. Statist.* **29** 94–123.

Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankhyā Ser. A* **31** 23–36.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`

Reiss, R.-D. (1973). On the measurability and consistency of maximum likelihood estimates for unimodal densities. *Ann. Statist.* **1** 888–901.

Reiss, R.-D. (1976). On minimum distance estimators for unimodal densities. *Metrika* **23** 7–14.

Robertson, T., Wright, F. and Dykstra, R. (1988). *Order Restricted Statistical Inference.* John Wiley & Sons Ltd.

Sen, B. and Meyer, M. C. (2013). Testing against a linear regression model using ideas from shape-restricted estimation. *Preprint.* ArXiv:1311.6849.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics, Springer-Verlag.

Wegman, E. J. (1968). *On estimating a unimodal density.* ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The University of Iowa.

Wegman, E. J. (1969). A note on estimating a unimodal density. *Ann. Math. Statist.* **40** 1661–1667.

Wegman, E. J. (1970a). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41** 457–471.

Wegman, E. J. (1970b). Maximum likelihood estimation of a unimodal density. II. *Ann. Math. Statist.* **41** 2169–2174.

Supplementary Material

Appendices

A Some further empirical coverage results

Table 3: Empirical coverage probabilities for the proposed confidence bands with $\alpha = 0.05$.

	β	R	$n = 100$	$n = 1000$	$n = 5000$
mixture of uniforms	0	S_n	**0.972**	**0.963**	**0.959**
		$\text{supp}(p_0)$	0	**0.953**	**0.959**
	0.25	S_n	**0.991**	**0.971**	**0.970**
		$\text{supp}(p_0)$	0	**0.961**	**0.970**
	0.5	S_n	**0.959**	**0.953**	**0.991**
		$\text{supp}(p_0)$	0	**0.944**	**0.991**
double logarithmic $p = 0.9$	0	S_n	**0.956**	**0.949**	**0.949**
		$p_0(R) \approx 0.95$	0.001	0.922	**0.949**
	0.25	S_n	0.970	**0.950**	**0.948**
		$p_0(R) \approx 0.95$	0.001	0.922	**0.948**
	0.5	S_n	0.980	0.989	0.989
		$p_0(R) \approx 0.95$	0.001	**0.959**	0.990

Let $R \subseteq S_0$. The results in Table 3 give the empirical coverage *on the set* R as indicated in the third column. That is, we report the proportion of times that

$$\widehat{c}_{n,l}(s) \quad \leq \quad p_0(s) \quad \leq \quad \widehat{c}_{n,l}(s), \quad \text{for all } s \in R$$

was observed. The confidence bands are optimized for $R = \text{supp}(\widehat{p}_n)$, and this allows us to compare the behavior for other choices of R.

The results of Table 3 clearly show the cost of only defining the confidence bands on $\text{supp}(\widehat{p}_n) = S_n$ in the simulations. For larger sample sizes, this cost decreases. However, for small sample sizes, the undercoverage is drastically big, simply because S_n does not cover the set R yet. This issue aside, we find that the confidence bands perform rather well. In the double logarithmic setting for $\beta < 0.5$, we expect to obtain asymptotically correct coverage bands, and hence, empirical coverage probabilities not statistically different from 0.95 are shown in bold. In all uniform mixture scenarios, we expect an asymptotically conservative result; that is, the asymptotic coverage should is expected to be greater than 0.95. In the table, empirical coverage probabilities not statistically smaller than 0.95 are shown in bold (for the mixture of uniforms case).

B Proofs and technical details

B.1 Note on finding the MLE

In several proofs, we make use of the following idea (a well known practice in shape-constrained estimation problems):

To compute $\widehat{p}_n|_\kappa$, we first relax consideration over pmfs to positive sequences, $\mathcal{U}|_\kappa(S_n)$, by changing the criterion function L_n to

$$\Phi_n(p) \quad = \quad L_n(p) - \sum_{j=0}^{J-1} p(z_j) \quad = \quad L_n(p) - \sum_{z \in S_n} p(z). \qquad \text{(B-18)}$$

This is possible because the two maximization problems are equivalent. To see this, note that if p is a positive sequence with support S_n maximizes Φ_n, then for all $c \in \mathbb{R}$ with $|c|$ very small

$$0 = \lim_{c \to 0} \frac{\Phi_n(p + cp) - \Phi_n(p)}{c} = 1 - \sum_{z \in S_n} p(z)$$

implying that p is necessarily a pmf. In the sequel, we denote by $\mathcal{U}(S_n)$ the space of positive unimodal sequences with support S_n.

B.2 Proofs from Section 4

Proof of Theorem 4.1. We first recall that

$$\rho(p|p_0) \quad \geq \quad 0, \qquad \text{(B-19)}$$

with equality if and only if $p = p_0$ (P_0 a.s.). This result is often referred to as Gibbs' inequality. We now proceed progressively in steps. We first assume that $|S_0| \geq 2$.

1. Let Q denote the cdf of any discrete distribution on $\mathbb{N}$ and $\widehat{Q}$ denote its least concave majorant (on $\mathbb{N}$). Let $\widehat{q}$ denote the pmf associated with $\widehat{Q}$. We first claim that $\widehat{q}$ is such that

$$\int \log \frac{\widehat{q}}{p} dQ \quad \geq \quad 0, \qquad \text{for all decreasing pmf } p.$$

This follows from the results of Patilea (1997, 2001) for decreasing densities as follows: Let $F_0(z) = Q(z-1)$ denote a cdf on $\mathbb{R}_+$, and let $\widehat{F_0}$ denote its least concave majorant on $\mathbb{R}_+$ (LCM), with associated pdf $\widehat{f_0}$. Then, from Patilea (1997, 2001) it follows that $\widehat{f_0}$ satisfies

$$\int \log \frac{\widehat{f_0}}{f} dF_0 \quad \geq \quad 0,$$

for all decreasing densities f, and hence also for all decreasing densities with the form $f(x) = \int_0^\infty \theta^{-1} 1_{[0,\theta]}(x) d\mu(\theta)$, where μ is discrete with mass only at $\mathbb{Z}_+$. In other words, any f which is piecewise constant, with points of jump occurring possibly only at $\mathbb{Z}_+$. For such densities f, let $p(z) = \int_z^{z+1} f(x)dx = f(z+1), z \in \mathbb{N}$. In addition, note that $\widehat{q}(z) = \int_z^{z+1} \widehat{f}_0(x)dx = \widehat{f}_0(z+1)$, from the definition of $\widehat{Q}$ and $\widehat{F}_0$. Then we have that

$$\begin{aligned} \int \log \frac{\widehat{q}}{p} dQ &= \sum_{z\geq 0} \log \frac{\widehat{q}(z)}{p(z)} q(z) \\ &= \sum_{z\geq 0} \log \frac{\widehat{f}_0(z+1)}{f(z+1)} \{F_0(z+1) - F_0(z)\} \\ &= \int \log \frac{\widehat{f}_0}{f} dF_0 \quad \geq \quad 0, \end{aligned}$$

and the result follows.

2. Next, let $\alpha = \sum_{i\geq\kappa} p_0(s_i)$, $q_1(i) = \alpha^{-1} p_0(s_{i+\kappa})$, and $q_2(i) = (1-\alpha)^{-1} p_0(s_{\kappa-1-i})$. Both q_1 and q_2 are pmfs on $\mathbb{N}$ and we apply step one above to find their $\widehat{q}_1, \widehat{q}_2$. Define $\widehat{p}_0|_\kappa(s_i) = \alpha \widehat{q}_1(i-\kappa), i \geq \kappa$ and $\widehat{p}_0|_\kappa(s_i) = (1-\alpha)\widehat{q}_2(\kappa-1-i)$ for $i \leq \kappa - 1$. Then clearly $\widehat{p}_0|_\kappa \in \mathcal{U}^1|_\kappa(S_0)$. Furthermore, for any $p \in \mathcal{U}^1|_k(S_0)$

$$\begin{aligned} \int \log \frac{\widehat{p}_0|_\kappa}{p} dP_0 &= \sum_{i\leq\kappa-1} \log \frac{\widehat{p}_0|_\kappa(s_i)}{p(s_i)} p_0(s_i) + \sum_{i\geq\kappa} \log \frac{\widehat{p}_0|_\kappa(s_i)}{p(s_i)} p_0(s_i) \\ &= \sum_{i\leq\kappa-1} \log \frac{(1-\alpha)\widehat{q}_2(\kappa-1-i)}{p(s_i)} p_0(s_i) \\ &\qquad + \sum_{i\geq\kappa} \log \frac{\alpha \widehat{q}_1(i-\kappa)}{p(s_i)} p_0(s_i) \\ &= (1-\alpha) \sum_{i\leq\kappa-1} \log \frac{(1-\alpha)\widehat{q}_2(\kappa-1-i)}{p(s_i)} \frac{p_0(s_i)}{1-\alpha} \\ &\qquad + \alpha \sum_{i\geq\kappa} \log \frac{\alpha \widehat{q}_1(i-\kappa)}{p(s_i)} \frac{p_0(s_i)}{\alpha} \\ &= (1-\alpha) \sum_{z\geq 0} \log \frac{\widehat{q}_2(z)}{p_2(z)} q_2(z) + \alpha \sum_{z\geq 0} \log \frac{\widehat{q}_1(z)}{p_1(z)} q_1(z), \end{aligned}$$

where $p_2(z) = (1-\alpha)^{-1} p(s_{\kappa-1-z})$ and $p_1(z) = \alpha^{-1} p(s_{\kappa+z})$. Now, let c_1 denote the constant such that $c_1 \sum_{z\geq 0} p_1(z) = 1$, and let $\widetilde{p}_1 = c_1 p_1$ (and similarly for p_2). Let $\beta = \sum_{z\geq 0} p(s_{\kappa+z})$. Then, we have that the above is equal to

$$\begin{aligned}
&(1-\alpha)\sum_{z\geq 0}\log\frac{\widehat{q_2}(z)}{p_2(z)}q_2(z)+\alpha\sum_{z\geq 0}\log\frac{\widehat{q_1}(z)}{p_1(z)}q_1(z)\\
&\quad= \quad (1-\alpha)\log c_2+\alpha\log c_1+(1-\alpha)\sum_{z\geq 0}\log\frac{\widehat{q_2}}{\widetilde{p_2}(z)}q_2(z)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad +\alpha\sum_{z\geq 0}\log\frac{\widehat{q_1}}{\widetilde{p_1}(z)}q_1(z)\\
&\quad\geq \quad (1-\alpha)\log c_2+\alpha\log c_1 \;=\; (1-\alpha)\log\frac{(1-\alpha)}{(1-\beta)}+\alpha\log\frac{\alpha}{\beta}\;\geq\; 0,
\end{aligned}$$

where the last inequality follows from the Gibbs' inequality in (B - 19) applied to two Bernoulli distributions with success probabilities α and β respectively. It follows that

$$\int\log\frac{\widehat{p_0}|_\kappa}{p}dP_0\geq 0$$

for any $p\in\mathcal{U}^1|_\kappa(S_0)$. We have therefore proved existence of $\widehat{p_0}|_\kappa$.

3. Finally, we prove that $\widehat{p_0}|_\kappa$ as defined above is the unique solution to (4.11) in the two cases stated in the proposition.

   • Suppose that $p_0\in\mathcal{U}^1|_\kappa(S_0)$. Then, by Gibbs' inequality, we have that

   $$\int\log\frac{p_0}{p}dP_0 \quad\geq\quad 0,\quad \forall p\in\mathcal{U}^1|_\kappa(S_0).$$

   Suppose then that $\widehat{p_0}$ is another candidate for the KL projection, as above. Then we would have that

   $$\int\log\frac{p_0}{\widehat{p_0}}dP_0 \quad\geq\quad 0 \qquad\text{and also}\qquad \int\log\frac{\widehat{p_0}}{p_0}dP_0 \quad\geq\quad 0.$$

   But this implies that

   $$\int\log\frac{p_0}{\widehat{p_0}}dP_0 \quad=\quad 0,$$

   and (again by Gibbs' inequality) it follows that $\widehat{p_0}=p_0$, P_0 a.s..

   • Suppose that $p_0\notin\mathcal{U}^1|_\kappa(S_0)$ with $\sum_{j\neq 0}\log|j|p_0(s_j)<\infty$. Then, by Proposition 4.3, we have that $\sup_{p\in\mathcal{U}^1|_\kappa(S_0)}\int\log p\,dP_0\in(-\infty,0]$. Hence, (4.11) is equivalent to

   $$\widehat{p_0}=\mathrm{argmax}_{p\in\mathcal{U}^1|_\kappa(S_0)}\int\log p\,dP_0.$$

By the strict concavity of log, we have that $\log(\alpha a + (1-\alpha)b) \geq \alpha \log a + (1-\alpha)\log b$, with equality iff $a = b$. Suppose that $\widehat{p}_1$ and $\widehat{p}_2$ are two different pmfs at which the cross entropy achieves its maximum. Then, by convexity of $\mathcal{U}^1|_\kappa(S_0)$, $\widehat{p}_0 = \alpha\widehat{p}_1 + (1-\alpha)\widehat{p}_2$ is also in $\mathcal{U}^1|_\kappa(S_0)$, and hence

$$\begin{aligned}
\int \log \widehat{p}_0 \, dP_0 &= \int \log\left\{\alpha\widehat{p}_1 + (1-\alpha)\widehat{p}_2\right\} dP_0 \\
&> \alpha \int \log \widehat{p}_1 dP_0 + (1-\alpha)\int \log \widehat{p}_2 dP_0 \\
&= \operatorname{argmax}_{p \in \mathcal{U}^1|_\kappa(S_0)} \int \log p \, dP_0
\end{aligned}$$

which yields a contradiction. Therefore, we must have that $\widehat{p}_1 = \widehat{p}_2$, P_0-almost surely. But this implies that on the set $\{s \in S_0 : p_0(s) > 0\}$, $\widehat{p}_1$ and $\widehat{p}_2$ must both be equal to the slope of the greatest convex minorant (GCM) of the cumulative sum of $p_0(s_i)$ to the left of $s_{\kappa-1}$ and the slope of its LCM to the right of s_κ. Since the latter has the same support as p_0, we conclude that uniqueness has to hold everywhere.

Lastly, suppose that $|S_0| = 1$. Then $p_0 \in \mathcal{U}^1|_\kappa(S_0)$ must be unimodal with $S_0 = \{s_\kappa\}$, and $\widehat{p}_0 = p_0$. The same proof as the first part of point three above applies. □

Lemma B.1. *Suppose that* $\sum_{j\neq 0} \log|j|\, p_0(s_j) < \infty$. *Then for each* $\kappa \in \mathbb{Z}$ *such that* $s_\kappa \in S_0$, *there exists a* $q \in \mathcal{U}^1|_\kappa(S_0)$ *such that* $\int \log q \, dP_0(x) \in (-\infty, 0]$.

Proof. Fix $\kappa \in \mathbb{Z}$ with $\kappa \neq 0$. Then

$$\begin{aligned}
\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j) &= \sum_{j\notin\{\kappa,0\}} \log |j| \frac{|j-\kappa|}{|j|} p_0(s_j) + \log|\kappa|\, p_0(s_0) \\
&= \sum_{j\neq 0,\kappa} \log|j|\, p_0(s_j) + \sum_{j\notin\{\kappa,0\}} \log\frac{|j-\kappa|}{|j|} p_0(s_j) + \log|\kappa|\, p_0(s_0) \\
&\leq \sum_{j\neq 0} \log|j|\, p_0(s_j) + \sum_{j\notin\{\kappa,0\}} \log\frac{|j-\kappa|}{|j|} p_0(s_j) + \log|\kappa|.
\end{aligned}$$

Now, since $\lim_{|j|\to\infty} \log\left(|j-\kappa|/|j|\right) = 0$, and by the assumption of the lemma, all three terms above are finite, and it follows that $\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j) < \infty$ for all $\kappa \in \mathbb{Z}$.

Define a pmf q with support S_0 as

$$q(s_j) \propto \begin{cases} \frac{1}{|j-\kappa|\log^2|j-\kappa|} & j \neq \kappa-1, \kappa, \kappa+1, \\ \frac{1}{2\log^2 2} & j = \kappa-1, \kappa, \kappa+1, \end{cases} \qquad \text{(B-20)}$$

for $s_j \in S_0$. Since $\int_2^\infty (x\log^2 x)^{-1} dx = 1/(\log 2)$, there exists a normalizing constant for $q \in \mathcal{U}^1|_\kappa(S_0)$. It remains to calculate its entropy. That is,

$$\sum_j \log q(s_j) p_0(s_j) = D - \sum_{|j-\kappa|\geq 2} \log|j-\kappa| p_0(s_j) - 2\sum_{|j-\kappa|\geq 2} \log\log|j-\kappa| p_0(s_j),$$

where D is some finite constant. The second term is also finite by the first part of this proof. For the last term we have that

$$0 \;\leq\; \sum_{|j-\kappa|\geq e} \log\log|j-\kappa| p_0(s_j) \;\leq\; \sum_{|j-\kappa|\geq e} \log|j-\kappa| p_0(s_j),$$

and hence this term is also finite. The result follows. □

Proof of Theorem 4.2. The first point can be shown using Gibbs' inequality as done above in the proof of Theorem 4.1. To prove the second point, we first note that by Proposition 4.3, under the assumptions of the theorem, (4.9) and (4.10) are equivalent. Therefore, to prove that (4.9) holds, it is sufficient to show that (4.10) holds, for some $\widehat{p}_0$. By Lemma B.1, for each $\kappa \in \mathbb{Z}$ there exists a $q \in \mathcal{U}^1|_\kappa(S_0)$ such that $\int \log q dP_0 > -\infty$. Therefore, each $\int \log \widehat{p}_0|_\kappa(x)\, dP_0(x) > -\infty$ (although this bound is not uniform in κ). Next, by Lemma C.1, we have that

$$\begin{aligned} \int \log \widehat{p}_0|_\kappa(x)\, dP_0(x) &= \sum_j \log \widehat{p}_0|_\kappa(s_j) p_0(s_j) \\ &\leq -\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j) \\ &\leq -\log|\kappa - m|\, p_0(s_m), \end{aligned}$$

for some fixed m such that $s_m \in S_0$. Letting $\kappa \to \pm\infty$, it follows that the maximum cannot be attained for large values of $|\kappa|$, and hence the supremum of $\int \log \widehat{p}_0|_\kappa(x)\, dP_0(x)$ can be found by considering a finite collection of values of κ. This proves existence of a maximizer $\widehat{p}_0 \in \mathcal{U}^1(S_0)$ (and also that $\{\widehat{p}_0\}$ is a finite set). □

Proof of Proposition 4.3. We first show that if $\sum_{j\neq 0} \log|j| dP_0 = \infty$, then $\int \log p dP_0 = -\infty$, for any unimodal p. This follows since, if p is unimodal, then $p \in \mathcal{U}^1|_\kappa(S_0)$ for some $\kappa \in \mathbb{Z}$. Hence, by Lemma C.1, we have that

$$\begin{aligned} \int \log p\, dP_0 &\leq \sum_j \log \min\left(1, |j-\kappa|^{-1}\right) p_0(s_j) \\ &= -\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j). \end{aligned}$$

Now, if $\kappa = 0$, then $\int \log p\, dP_0 \leq -\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j) = \sum_{j\neq 0} \log|j|\, p_0(s_j)$. If $\kappa \neq 0$, then

$$\begin{aligned} -\sum_{j\neq\kappa} \log|j-\kappa|\, p_0(s_j) &= -\sum_{j\notin\{\kappa,0\}} \log|j| \frac{|j-\kappa|}{|j|}\, p_0(s_j) - \log|\kappa|\, p_0(s_0) \\ &= -\sum_{j\neq 0} \log|j|\, p_0(s_j) - \sum_{j\notin\{\kappa,0\}} \log \frac{|j-\kappa|}{|j|} p_0(s_j) - \log|\kappa|\, p_0(s_0) + \log|\kappa| p_0(s_\kappa) \\ &\leq -\sum_{j\neq 0} \log|j|\, p_0(s_j) - \sum_{j\notin\{\kappa,0\}} \log \frac{|j-\kappa|}{|j|} p_0(s_j) + \log|\kappa|. \end{aligned}$$

Since $\lim_{|j|\to\infty} \log(|j-\kappa|/|j|) = 0$, there exists an integer $J > 0$ such that for all $|j| > J$

$$\log \frac{|j-\kappa|}{|j|} \in \left[-\frac{1}{2}, \frac{1}{2}\right].$$

Then,

$$\begin{aligned}
&\sum_{j\notin\{\kappa,0\}} \log \frac{|j-\kappa|}{|j|} p_0(s_j) \\
&\qquad = \sum_{j\notin\{\kappa,0\},|j|\le J} \log \frac{|j-\kappa|}{|j|} p_0(s_j) + \sum_{j\notin\{\kappa,0\},|j|>J} \log \frac{|j-\kappa|}{|j|} p_0(s_j) \\
&\qquad \ge \sum_{j\notin\{\kappa,0\},|j|\le J} \log \frac{|j-\kappa|}{|j|} p_0(s_j) - \frac{1}{2} \sum_{j\notin\{\kappa,0\},|j|>J} p_0(s_j) \\
&\qquad \ge \sum_{j\notin\{\kappa,0\},|j|\le J} \log \frac{|j-\kappa|}{|j|} p_0(s_j) - \frac{1}{2} = -C,
\end{aligned}$$

for some finite constant C. Therefore,

$$\int \log p \, dP_0 \quad \le \quad -\sum_{j\ne 0} \log|j| p_0(s_j) + C + \log|\kappa|,$$

and the first part of the claim follows (noting that since $\kappa \in \mathbb{Z}$ and $\kappa \ne 0$, then $\log|\kappa| < \infty$). The second part of the claim follows immediately from Lemma B.1. □

B.2.1 Proof of Theorem 4.4

We start by showing the following lemma.

Lemma B.2. *Suppose that* $\sum_{i\ne 0} \log|i| p_0(s_i) < \infty$. *Let* $\widehat{\mathcal{M}}_n$ *be the modal region of* $\widehat{p}_n$. *Then, we can find* $M > 0$ *sufficiently large, such that with probability one there exists an integer* $n_0 > 0$ *such that*

$$\sup_{n\ge n_0} \max_{\kappa\in\widehat{\mathcal{M}}_n} |\kappa| \le M + 1.$$

Proof. Fix $\varepsilon_1 \in (0, p_0(s_0)/4)$, and define the event that $A_n^c = \{\sup|\mathbb{F}_n - F_0| \le \varepsilon_1\}$. By the Dvoretzky-Kiefer-Wolfowitz inequality the probability of A_n is at most $2e^{-2n\varepsilon_1^2}$. Applying Lemma C.1, we have that

$$\begin{aligned}
\int \log \widehat{p}_n|_\kappa d\mathbb{F}_n \quad &\le \quad -\sum_{i\ne\kappa} \log|i-\kappa| \overline{p}_n(s_i) \\
&\le \quad -\log|\kappa| \overline{p}_n(s_0) \quad \le \quad -\log|\kappa|(p_0(s_0) - 2\varepsilon_1) \\
&\le \quad -\log|\kappa| p_0(s_0)/2 \quad \le \quad -\log M \, p_0(s_0)/2,
\end{aligned}$$

if $|\kappa| > M$. Let B_n denote the event that $\int \log \widehat{p}_n|_\kappa d\mathbb{F}_n > -\log M \, p_0(s_0)/2$, whenever $|\kappa| > M$. By the above, we have that $B_n \subset A_n$. Since $P(A_n)$ is summable, the Borel-Cantelli lemma implies that $P(B_n \text{ i.o.}) = 0$. Thus, we have shown that, with probability one, there exists an integer n_1 such that for all $n \geq n_1$

$$\int \log \widehat{p}_n|_\kappa d\mathbb{F}_n \quad \leq \quad -\log M\, p_0(s_0)/2, \quad \forall |\kappa| > M.$$

Without loss of generality, we can assume that $S_0 = \{s_i, i \in K\}$ with $K = \mathbb{Z}$. Next, define q as in (B - 20), and note that here we have

$$\sum_i |\log q(s_i)| p_0(s_i) < \infty,$$

using similar arguments to those used in the proof of Proposition 4.3. Recall that the pmf $q \in \mathcal{U}^1|_{\kappa=0}(S_0)$. Fix $\varepsilon_2 > 0$. By the strong law of large numbers, we can find with probability one an integer n_2 such that for all $n \geq n_2$

$$\int \log q\, d\mathbb{F}_n \quad \geq \quad \int \log q\, dF_0 - \varepsilon_2.$$

Since $\int \log q\, dF_0 \in \mathbb{R}$, we can furthermore choose ε_2 and M so that

$$\int \log q\, d\mathbb{F}_n \quad > \quad -\log M\, p_0(s_0)/2.$$

Thus, it follows that with probability one, there exists an n_0 (in fact, $n_0 = \max\{n_1, n_2\}$), such that

$$\int \log q\, d\mathbb{F}_n \quad > \quad \int \log \widehat{p}_n|_\kappa d\mathbb{F}_n$$

for all $|\kappa| > M$. But this implies that $\widehat{p}_n|_\kappa$ cannot be equal to the MLE $\widehat{p}_n$ when $|\kappa| > M$, proving the result. □

Proof of Theorem 4.4. We want to show that $\widehat{p}_n \to \widehat{p}_0$. First, we recall that pointwise convergence and convergence in $\ell_k, 1 \leq k \leq \infty$ and Hellinger distance h are all equivalent for sequences of pmfs. This follows for example from Lemma C.2 in the on-line supporting material of Balabdaoui et al. (2013). We also recall that a collection of probability measures is tight if, for all $\varepsilon > 0$, there exists a compact set $K = K(\varepsilon)$, such that for all measures μ in the collection, we have $\mu(K^c) < \varepsilon$. Let $\widehat{P}_n$ denote the measure induced by $\widehat{p}_n$. We first claim that $\{\widehat{P}_n\}_{n\geq 1}$ is tight with probability one. Fix $\varepsilon > 0$. Then, by the Glivenko-Cantelli theorem, we can find with probability one an integer $n_1 > 0$ such that for all $n \geq n_1$, $\sup_{s \in S_0} |\mathbb{F}_n(s) - F_0(s)| < \varepsilon/6$. Also, by definition of the cdf, there exists a constant $M_0 > 0$ such that for all $M \geq M_0$,

$$1 - F_0(M) + F_0(-M-1) < \varepsilon/6.$$

Note that this implies that we have with probability one

$$\begin{aligned}
1 - \mathbb{F}_n(M) + \mathbb{F}_n(-M-1) & \\
&= 1 - F_0(M) + F_0(-M-1) \\
&\quad + \{F_0(M) - \mathbb{F}_n(M)\} + \{\mathbb{F}_n(-M-1) - F_0(-M-1)\} \\
&< \varepsilon/2,
\end{aligned}$$

for all $n \geq n_1$ and all $M \geq M_0$.

Next, let $\widehat{\kappa}_n$ be such that $\widehat{p}_n \in \mathcal{U}^1|_{\widehat{\kappa}_n}(S_0)$. Then, by the result of Lemma B.2, with probability one, there exist $M > M_0$ and an integer $n_2 > 0$ such that for all $n \geq n_2$, $\sup_{n \geq n_2} |\widehat{\kappa}_n| \leq M$. On this event, we have that

$$\begin{aligned}
\widehat{P}_n\left([-M, M]^c\right) &= \sum_{z \geq M+1} \widehat{p}_n(z) + \sum_{z \leq -M-1} \widehat{p}_n(z) \\
&\leq \sum_{z \geq M+1} \overline{p}_n(z) + \sum_{z \leq -M-1} \overline{p}_n(z) \\
&= 1 - \mathbb{F}_n(M) + \mathbb{F}_n(-M-1) < \varepsilon/2
\end{aligned}$$

where the inequality in the second line follows from Proposition C.2. We have therefore shown that there exists a sufficiently large $n_0 = \max\{n_1, n_2\}$ such that $\{\widehat{P}_n\}_{n \geq n_0}$ is tight. Since any finite collection of distributions is also tight, it follows that $\{\widehat{P}_n\}_{n \geq 1}$ is tight, with probability one.

Since $\{\widehat{P}_n\}$ is tight, it is also sequentially compact. Thus, let $\{\widehat{P}_{n_k}\}$ denote a weakly convergent subsequence, which, for convenience, we continue to denote as $\{\widehat{P}_n\}$. The Portmanteau theorem then implies that the associated pmf $\widehat{p}_n(s_i)$ converges for all $s_i \in S_0$ (since $(s - \delta, s + \delta)$ are continuity sets for appropriate choice of δ), and we let $\widetilde{p}$ denote the limiting pmf. To complete the proof, we need only show that $\widetilde{p}$ is an element of $\{\widehat{p}_0\}$. Note that convergence in the set metric then follows because $\{\widehat{p}_0\}$ is necessarily a finite set.

Now, since we maximize the criterion function $\int \log p \, d\mathbb{F}_n - \sum_{z \in S_n} p(z)$ (B-18) over positive and unimodal sequences and since $\sum_{z \in S_n} \widehat{p}_n(z) = 1$, we can write

$$\begin{aligned}
\sum \log \widehat{p}_0(z_j)\, \overline{p}_n(z_j) - \sum \widehat{p}_0(z_j) &\leq \sum \log \widehat{p}_n(z_j)\, \overline{p}_n(z_j) - 1 \\
&\leq \sum \log(b + \widehat{p}_n(z_j))\, \overline{p}_n(z_j) - 1,
\end{aligned}$$

for $b > 0$. Re-arranging the terms above, this yields

$$\begin{aligned}
0 &\leq \sum \log(b + \widehat{p}_n(z_j))\, \overline{p}_n(z_j) - \sum \log \widehat{p}_0(z_j)\, \overline{p}_n(z_j) + \sum \widehat{p}_0(z_j) - 1 \\
&\leq \sum \log(b + \widehat{p}_n(z_j))\, \overline{p}_n(z_j) - \sum \log \widehat{p}_0(z_j)\, \overline{p}_n(z_j),
\end{aligned}$$

where the last inequality follows since $\sum \widehat{p}_0(z_j) \leq 1$. Finally, because $\overline{p}_n$ puts all of its mass only on the z_j, we can re-write the latter as

$$0 \leq \sum \log(b + \widehat{p}_n(s_j))\, \overline{p}_n(s_j) - \sum \log \widehat{p}_0(s_j)\, \overline{p}_n(s_j).$$

On the other hand, we have that

$$\begin{aligned}
&\sum \log(b+\widehat{p}_n(s_j))\,\overline{p}_n(s_j) - \sum \log \widehat{p}_0(s_j)\,\overline{p}_n(s_j)\\
&= \sum \log(b+\widehat{p}_n(s_j))\,(\overline{p}_n(s_j)-p_0(s_j)) + \sum \log \widehat{p}_0(s_j)\,(p_0(s_j)-\overline{p}_n(s_j))\\
&\quad + \sum \log\left(\frac{b+\widehat{p}_n(s_j)}{b+\widehat{p}_0(s_j)}\right) p_0(s_j) + \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{\widehat{p}_0(s_j)}\right) p_0(s_j).
\end{aligned} \tag{B-21}$$

Next we get rid of the first two terms on the right-hand side. First, using summation by parts,

$$\begin{aligned}
&\sum \log\left(b+\widehat{p}_n(s_j)\right)\left(\overline{p}_n(s_j)-p_0(s_j)\right)\\
&\qquad = \sum (\mathbb{F}_n(s_j)-F_0(s_j))\left[\log\left(b+\widehat{p}_n(s_j)\right) - \log\left(b+\widehat{p}_n(s_{j-1})\right)\right].
\end{aligned}$$

Now, we know that $\widehat{p}_n = \widehat{p}_n|_\kappa$ for some κ. Then,

$$\begin{aligned}
&\left|\sum \log\left(b+\widehat{p}_n(s_j)\right)\left(\overline{p}_n(s_j)-p_0(s_j)\right)\right|\\
&\quad \le \ \sup|\mathbb{F}_n(s_j)-F(s_j)| \left\{ \sum_{j\le\kappa-1} \left[\log\left(b+\widehat{p}_n(s_j)\right) - \log\left(b+\widehat{p}_n(s_{j-1})\right)\right] \right.\\
&\qquad\qquad + \left|\log\left(b+\widehat{p}_n(s_\kappa)\right) - \log\left(b+\widehat{p}_n(s_{\kappa-1})\right)\right|\\
&\qquad\qquad \left. + \sum_{j\ge\kappa+1} \left[\log\left(b+\widehat{p}_n(s_j)\right) - \log\left(b+\widehat{p}_n(s_{j-1})\right)\right] \right\}\\
&\quad \le \ 4|\log(b+\max_j \widehat{p}_n(s_j))| \sup|\mathbb{F}_n(s_j)-F_0(s_j)|\\
&\quad \le \ 4\max\{\log(1+b), |\log(b)|\} \sup|\mathbb{F}_n(s_j)-F_0(s_j)|,
\end{aligned}$$

which converges to zero. The law of large numbers shows that the second term also converges to zero. This follows because $\sup_{p\in\mathcal{U}^1(S_0)} \int \log p\, dP_0 > -\infty$, which implies that $\sum |\log \widehat{p}_0(s_j)| p_0(s_j) < \infty$. Therefore, rearranging (B-21), we find that

$$\limsup_n \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{b+\widehat{p}_n(s_j)}\right) p_0(s_j) \quad \le \quad \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{\widehat{p}_0(s_j)}\right) p_0(s_j).$$

Now, letting $b \to 0$, we have by Fatou's lemma that

$$\limsup_{b\to 0} \limsup_n \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{b+\widehat{p}_n(s_j)}\right) p_0(s_j) \quad \le \quad 0.$$

Next, we take the limits on the right-hand side. First, by the dominated convergence theorem

$$\limsup_n \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{b+\widehat{p}_n(s_j)}\right) p_0(s_j) \quad = \quad \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{b+\widetilde{p}(s_j)}\right) p_0(s_j),$$

since $|\log((b+\widehat{p}_0)/(b+\widehat{p}_n))| \leq 2\max\{\log(b+1), |\log b|\}$. Next, we want to show that

$$\lim_{b\downarrow 0} \sum \log\left(\frac{b+\widehat{p}_0(s_j)}{b+\widetilde{p}(s_j)}\right) p_0(s_j) \;=\; \sum \log\left(\frac{\widehat{p}_0(s_j)}{\widetilde{p}(s_j)}\right) p_0(s_j). \quad \text{(B - 22)}$$

To do this, consider both pieces separately. First, $\log(b+\widehat{p}_0(z))$ is decreasing in b and bounded above by $\log 2$, and hence by the monotone convergence theorem we have that

$$\lim_{b} \sum \log\left(b+\widehat{p}_0(s_j)\right) p_0(s_j) \;=\; \sum \log \widehat{p}_0(s_j)\, p_0(s_j).$$

Similarly $-\log(b+\widetilde{p}(s_j))$ is increasing as b decreases, and bounded below by $-\log 2$. Therefore also,

$$\lim_{b} \sum \log\left(b+\widetilde{p}(s_j)\right) p_0(s_j) \;=\; \sum \log \widetilde{p}(s_j)\, p_0(s_j).$$

Note that $\int \log p\, dP_0$ is always finite for any unimodal p, and therefore we may subtract the last two lines above to yield (B - 22). We have thus shown that

$$\sum \log\left(\frac{\widehat{p}_0(s_j)}{\widetilde{p}(s_j)}\right) p_0(s_j) \;\leq\; 0.$$

Rearranging, this gives

$$\sup_{p\in\mathcal{U}^1(S_0)} \int \log p\, dP_0 \;=\; \sum \log \widehat{p}_0(s_i)\, p_0(s_i) \;\leq\; \sum \log \widetilde{p}(s_i)\, p_0(s_i),$$

and hence $\widetilde{p} \in \{\widehat{p}_0\}$. □

Recall the definition of knots in (5.14) and the preceding paragraph.

Lemma B.3. *Suppose that $\sum_{j\neq 0} \log|j| p_0(s_j) < \infty$ and $|\{\widehat{p}_0\}| = 1$. Let $\tau \in \mathcal{T}$ be a knot point of p_0. Then, almost surely, there exists an n_0 such that for all $n \geq n_0$ we have that τ is also a knot of $\widehat{p}_n$.*

Proof. Without loss of generality, assume that $\tau = s_{k_0}$ and that $\widehat{p}_0(s_{k_0}) > \widehat{p}_0(s_{k_0-1})$. Then, from Theorem 4.4, we know that $\sup|\widehat{p}_0(s_j) - \widehat{p}_n(s_j)| < \varepsilon$, where $\varepsilon < (\widehat{p}_0(s_{k_0}) - \widehat{p}_0(s_{k_0-1}))/2$, for all sufficiently large n. Therefore,

$$\widehat{p}_n(s_{k_0}) \geq \widehat{p}_0(s_{k_0}) - \varepsilon \quad > \widehat{p}_0(s_{k_0-1}) + \varepsilon \geq \widehat{p}_n(s_{k_0-1}),$$

and the result follows. □

Proof of Corollary 4.6. Write $\mathcal{M} = \{s_{k_1}, \ldots, s_{k_2}\}$, and note that, by definition, we have that $\widehat{p}_0(s_{k_1}) > \widehat{p}_0(s_{k_1-1})$ and $\widehat{p}_0(s_{k_2}) > \widehat{p}_0(s_{k_2+1})$. From

Lemma B.3 and the ℓ_1 consistency results of Theorem 4.4, it follows that, with probability one, there exists an n_0 such that for all $n \geq n_0$,

$$\begin{array}{rcll} \widehat{p}_n(s_j) & \leq & \widehat{p}_n(s_{k_1-1}) < \widehat{p}_n(s_{k_1}), & j \leq k_1 - 1, \\ \widehat{p}_n(s_{k_2}) & > & \widehat{p}_n(s_{k_2+1}) \geq \widehat{p}_n(s_j), & j \geq k_2 + 1. \end{array}$$

This, of course, implies that the mode of $\widehat{p}_n$ must be in $\mathcal{M} = \{s_{k_1}, \dots, s_{k_2}\}$. □

Proof of Corollary 4.7. This is an immediate consequence of Theorem 4.4 and the inequality

$$|\widehat{F}_n(s_i) - \widehat{F}_0(s_i)| \leq \sum_j |\widehat{p}_n(s_j) - \widehat{p}_0(s_j)|.$$

□

B.3 Proof of Theorem 5.1

Let $\{\mathbb{W}_n(s), s \in S_n\} \equiv \{\sqrt{n}(\overline{p}_n(s) - p_0(s)), s \in S_n\}$, denote the empirical white noise process.

Proposition B.4. *Let $C = \left\{\cup_{j=1}^k I_j\right\} \cup \left\{\cup_{j=1}^k D_j\right\}$ with k finite and I_j, D_j defined as in* (5.13)*. Then, with probability one, there exists an integer $n_0 > 0$ such that for $n \geq n_0$*

$$\sqrt{n}(\widehat{p}_n - p_0)(s) = \varphi[\mathbb{W}_n](s), \text{ for all } s \in C.$$

Proof. By the strong law of large numbers, with probability one, we can find an integer $n_1 > 0$ such that for all $n \geq n_1$, $C \subset S_n$. Next, by Corollary 4.6, with probability one, we can find $n_2 \geq n_1$ such that for $n \geq n_2$ we have that $\widehat{\mathcal{M}}_n \subset \mathcal{M}$. This means that the $\widehat{p}_n$ is found as the minimizer in $\mathcal{U}^1|_\kappa(S_0)$ where $\kappa \in \mathcal{M}$. By Lemma B.3, again with probability one, we can find an $n_3 \geq n_2$ such that the knots $\tau_i^I, \tau_i^D, i = 0, \dots, k$ are also knots of $\widehat{p}_n$ for all $n \geq n_3$ (recall the definitions of the knots from (5.14) and the preceding paragraph). Therefore, by Lemma C.4, for all $n \geq n_3$ we have that for $1 \leq j \leq k$

$$\begin{array}{rcl} \widehat{p}_n(s) & = & \text{iso}[(\overline{p}_n)_{I_j}](s), s \in I_j, \\ \widehat{p}_n(s) & = & \text{anti}[(\overline{p}_n)_{D_j}](s), s \in D_j. \end{array}$$

That is, we have that $\widehat{p}_n(s) = \varphi[\overline{p}_n](s), s \in C$, for $n \geq n_3$. Since p_0 is constant on each I_j, D_j by definition, this implies that

$$\sqrt{n}(\widehat{p}_n - p_0)(s) = \varphi[\mathbb{W}_n](s), \text{ for all } s \in C,$$

see Lemma C.5. □

Lemma B.5. *Let $\mathbb{V}$ be a mean-zero Gaussian vector of dimension $d > 0$ with variance-covariance matrix Σ given by $cov(\mathbb{V}_i, \mathbb{V}_j) = d^{-1}\delta_{i=j} - d^{-2}$. Then* uni$[\mathbb{V}]$ *is unique with probability one.*

Proof. Suppose that $\widehat{\mathbb{V}}_1$ and $\widehat{\mathbb{V}}_2$ are two different solutions for the minimization problem. Our goal will be to show that $P(\widehat{\mathbb{V}}_1 \neq \widehat{\mathbb{V}}_2) = 0$. Since any minimizer of uni$(\mathbb{V})$ can be re-written as local averages of the original vector $\mathbb{V}$, it follows that we can find $d \times d$ matrices $\widehat{A}_1$ and $\widehat{A}_2$ such that $\widehat{\mathbb{V}}_1 = \widehat{A}_1\mathbb{V}$ and $\widehat{\mathbb{V}}_2 = \widehat{A}_2\mathbb{V}$, where $\widehat{A}_i, i = 1, 2$ can be written as

$$\widehat{A}_i = \begin{bmatrix} \widehat{A}_i^1 & 0 & 0 & \dots & 0 \\ 0 & \widehat{A}_i^2 & 0 & \dots & 0 \\ & & \dots & & \\ 0 & 0 & 0 & 0 & \widehat{A}_i^{m_i} \end{bmatrix},$$

with $\widehat{A}_i^j,\ 1 \leq j \leq m_i$, given by the $l_j \times l_j$ matrix

$$\widehat{A}_i^j = \frac{1}{l_j}\begin{bmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}.$$

Also, note that if $\widehat{\mathbb{V}}_i = \widehat{A}_i\mathbb{V}$ then

$$\|\widehat{\mathbb{V}}_i - \mathbb{V}\|_2^2 \quad = \quad \mathbb{V}^T(I - \widehat{A}_i)\mathbb{V}.$$

Finally, let $\mathcal{A}$ denote the set of all possible matrices $\widehat{A}_i$, and note that $|\mathcal{A}|$ is finite. Hence,

$$\begin{aligned} P\left(\widehat{\mathbb{V}}_1 \neq \widehat{\mathbb{V}}_2\right) &= P\left(\widehat{\mathbb{V}}_1 \neq \widehat{\mathbb{V}}_2, \mathbb{V}^T(I - \widehat{A}_1)\mathbb{V} = \mathbb{V}^T(I - \widehat{A}_2)\mathbb{V}\right) \\ &\leq \sum_{B_1, B_2, \in \mathcal{A}, B_1 \neq B_2} P\left(\mathbb{V}^T(I - B_1)\mathbb{V} = \mathbb{V}^T(I - B_2)\mathbb{V}\right) \\ &= \sum_{B_1, B_2, \in \mathcal{A}, B_1 \neq B_2} P\left(\mathbb{V}^T(B_1 - B_2)\mathbb{V} = 0\right). \end{aligned}$$

Let $S = \Sigma^{1/2}$ so that we can write $\mathbb{V} = SZ$ for $Z \sim \mathcal{N}_d(0, I)$. The matrix $S^T(B_1 - B_2)S$ is Hermitian, and therefore admits a spectral decomposition, which we write as $\Gamma\Lambda\Gamma^T$, where Γ is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p, -\lambda_{p+1}, \dots, -\lambda_d)$ with $\lambda_i \geq 0,\ 1 \leq i \leq d$. Note that since $B_1 \neq B_2$, there exists at least one index $i \in \{1, \dots, d\}$ such that $\lambda_i \neq 0$. It is also important to note that only $B \in \mathcal{A}$ with $m = 1$ yields $B\mathbb{V} = 0$. Finally, let $U = \Gamma^T Z$. Note that $U \sim \mathcal{N}(0, I)$. Then, we can write

$$\begin{aligned} P\left(\mathbb{V}^T(B_1 - B_2)\mathbb{V} = 0\right) &= P(Z^T\Gamma\Lambda\Gamma^T Z = 0) \\ &= P(U^T\Lambda U = 0) \\ &= P(\lambda_1 U_1^2 + \dots + \lambda_p U_p^2 = \lambda_{p+1}U_{p+1}^2 + \dots + \lambda_d U_d^2). \end{aligned}$$

Notice that in the last line at least one of the quantities on the left or right hand side is not equal to zero and that in such case it is a continuous random variable (in fact, each has a gamma distribution). Also, notice that the left hand side is independent of the right hand side. This shows that $P\left(\mathbb{V}^T(B_1 - B_2)\mathbb{V} = 0\right) = 0$, and the result follows. □

Proof of Theorem 5.1. The proof is divided into several main steps. We first address a slight technicality: the MLE $\widehat{p}_n$ is defined on S_n, while p_0 is defined on S_0. The results we prove here all "live" in the space of ℓ_k sequences defined on S_0. To make our results concrete, we therefore embed all sequences on S_n into sequences on S_0 by setting them equal to zero for $s \notin S_n$.

Below, we present the proof for $\ell_k(S_0)$ with $k = 2$ only. Convergence for $3 \le k \le \infty$ follows immediately, because $\|q\|_k \le \|q\|_2$ for $k \ge 2$ and $q \in \ell_2$ and hence $\|\cdot\|_k$ is a continuous mapping on $\ell_2(S_0)$ for $k \ge 2$.

1. We first show that $\mathbb{W}_n$ converges in $\ell_2(S_0)$ to the limit $\mathbb{W}$. This is essentially a well known result (cf. Jankowski and Wellner (2009, Theorem 3.1)), noting that for $s \notin S_n$, $\mathbb{W}_n(s) = -\sqrt{n}p_0(s) = \sqrt{n}(\overline{p}_n(s) - p_0(s))$ is still well-defined, since for $s \notin S_n, \overline{p}_n(s) = 0$.

2. We will next show that $\sqrt{n}(\widehat{p}_n - p_0) \Rightarrow \varphi[\mathbb{W}]$ in $\ell_2(S_0 \setminus \mathcal{M})$. That is, we consider the sequence *only* on the set $S_0 \setminus \mathcal{M}$. This result is proved in two sub-steps:

   (a) We first show that φ is continuous in $\ell_2(S_0 \setminus \mathcal{M})$. This, together with step one above implies that $\varphi[\mathbb{W}_n] \Rightarrow \varphi[\mathbb{W}]$ in $\ell_2(S_0 \setminus \mathcal{M})$.

   (b) The next step is to show that $\|\sqrt{n}(\widehat{p}_n - p_0) - \varphi[\mathbb{W}_n]\|_2^2 \overset{p}{\to} 0$ (where the ℓ_2 norm is calculated only on the support $S_0 \setminus \mathcal{M}$). In fact, we prove slightly stronger convergence (in expectation).

3. Finally, we will tackle convergence on the set $\mathcal{M}$. This follows essentially from the argmax continuous mapping theorem. Note that since $|\mathcal{M}|$ is necessarily finite, we also have convergence in $\ell_2(\mathcal{M})$ of the process on the set $\mathcal{M}$.

4. To put the two results together, note that the convergence in steps two and three can also be stated as joint convergence (and not *just* convergence of marginals). This holds because of the joint convergence of $\mathbb{W}_n$ in step one. From here the full result follows.

We now fill in the details in steps 2 and 3 above. To prove 2(a), consider a converging sequence $q_n \to q$ in $\ell_2(S_0 \setminus \mathcal{M})$ and fix $\varepsilon > 0$. Then we can find an integer n_0 and $K > 0$ large enough such that

$$\sup_{n \ge n_0} \sum_{|i| > K} q_n^2(s_i) < \varepsilon/6, \text{ and } \sum_{|i| > K} q^2(s_i) < \varepsilon/6.$$

Now, let $K_1 \le -K$ and $K_2 \ge K$ be such that $s_{K_1}, s_{K_2} \in \mathcal{T}$. We then have that

$$\begin{aligned}\sum_{s_i\notin\mathcal{M}}(\varphi[q_n](s_i)-\varphi[q](s_i))^2 \quad &\le \quad \sum_{i\in[K_1,K_2],s_i\notin\mathcal{M}}(\varphi[q_n](s_i)-\varphi[q](s_i))^2 \\ &\quad +2\sum_{i\notin[K_1,K_2]}\varphi[q_n]^2(s_i)+2\sum_{i\notin[K_1,K_2]}\varphi[q]^2(s_i).\end{aligned}$$

Now, by Lemma C.5 (choosing $p = q = 0$) we have that

$$\sum_{i\notin[K_1,K_2]}\varphi[q_n]^2(s_i) \le \sum_{i\notin[K_1,K_2]} q_n^2(s_i),$$

and similarly for q_n replaced with q. Also, by continuity of the operators iso and anti (Proposition C.6), we can choose an $n_1 \ge n_0$ such that for all $n \ge n_1$

$$\sum_{i\in[K_1,K_2],s_i\notin\mathcal{M}}(\varphi[q_n](s_i)-\varphi[q](s_i))^2 < \varepsilon/3.$$

It follows that for all $n \ge n_1$, we have that

$$\begin{aligned}\sum_{s_i\notin\mathcal{M}}(\varphi[q_n](s_i)-\varphi[q](s_i))^2 \quad &\le \quad \varepsilon/3+2\sum_{|i|\ge K}q_n^2(s_i)+2\sum_{|i|\ge K}q^2(s_i) \\ &\le \quad \varepsilon/3+4\varepsilon/6=\varepsilon.\end{aligned}$$

This shows that φ is continuous in $\ell_2(S_0 \setminus \mathcal{M})$.

To prove 2(b), we fix $\varepsilon > 0$ and pick K large enough so that $\sum_{|i|>K} p_0(s_i) < \varepsilon$. Now, let $K_1 \le -K$ and $K_2 \ge K$ be such that $s_{K_1}, s_{K_2} \in \mathcal{T}$. Let $\widehat{\mathbb{W}}_n(s) = \sqrt{n}(\widehat{p}_n - p_0)(s)$. Then

$$\begin{aligned}\sum_{s_i\notin\mathcal{M}}(\widehat{\mathbb{W}}_n-\varphi[\mathbb{W}_n])^2(s_i) \quad &\le \quad \sum_{i\in[K_1,K_2],s_i\notin\mathcal{M}}(\widehat{\mathbb{W}}_n(s_i)-\varphi[\mathbb{W}_n](s_i))^2 \\ &\quad +2\sum_{i\notin[K_1,K_2],s_i\in S_n}\widehat{\mathbb{W}}_n^2(s_i)+2\sum_{i\notin[K_1,K_2],s_i\in S_n}\mathbb{W}_n^2(s_i) \\ &\quad +4\sum_{i\notin[K_1,K_2],s_i\notin S_n}\mathbb{W}_n^2(s_i).\end{aligned}$$

Now, for n large enough, by Proposition B.4 we have that $\widehat{\mathbb{W}}_n(s_i) = \varphi[\mathbb{W}_n](s_i)$ for all $i \in [K_1, K_2], s_i \notin \mathcal{M}$. Also, $\sum_{i\notin[K_1,K_2],s_i\in S_n}\widehat{\mathbb{W}}_n^2(s_i) \le \sum_{i\notin[K_1,K_2],s_i\in S_n}\mathbb{W}_n^2(s_i)$ by Lemma C.5. We therefore have that for n sufficiently large

$$\|\widehat{\mathbb{W}}_n-\varphi[\mathbb{W}_n]\|_2^2 \quad \le \quad 4\sum_{|i|>K}\mathbb{W}_n^2(s_i).$$

$$\begin{aligned}0\le\overline{\lim}\,E\left[\|\widehat{\mathbb{W}}_n-\varphi[\mathbb{W}_n]\|_2^2\right] \quad &\le \quad E\left[\overline{\lim}\,\|\widehat{\mathbb{W}}_n-\varphi[\mathbb{W}_n]\|_2^2\right] \\ &\le \quad 4E\left[\overline{\lim}\sum_{|i|>K}\mathbb{W}_n^2(s_i)\right]\le 4E\left[\sum_{|i|>K}\mathbb{W}^2(s_i)\right] \\ &= \quad \sum_{|i|>K}p_0(s_i)<\varepsilon.\end{aligned}$$

Since ε was arbitrary, this proves the result.

Finally, we tackle step 3. We will do this by applying the argmax continuous mapping theorem, cf. van der Vaart and Wellner (1996, Theorem 3.2.2, page 286). Let $L_n(p)$ denote again the empirical log-likelihood, and recall that

$$\widehat{p}_n \quad = \quad \mathrm{argmax}_{\kappa \in S_n} L_n(\widehat{p}_n|_\kappa).$$

Now, by Lemma B.3 applied to τ_0^I and τ_0^D, and by Lemma C.4 we can also have that

$$\widehat{p}_n \quad = \quad \mathrm{argmax}_{\kappa \in \mathcal{M}} L_n(\widehat{p}_n|_\kappa),$$

and furthermore, each $\widehat{p}_n|_\kappa(s), s \in \mathcal{M}$ is determined by the LCM/GCM characterization only on $\mathcal{M}$. Let $d = |\mathcal{M}|$ and recall the definition of $\mathcal{U}_d$ from Section 2.2.2 as the space of unimodal vectors of length d. Also, let $\mathcal{U}_d^+ = \{u \in \mathcal{U}_d : u > 0\}$. For $s \in \mathcal{M}$, and for sufficiently large n, we have that

$$\begin{aligned}
\sqrt{n}(\widehat{p}_n - p_0) \quad &= \quad \sqrt{n}\left\{\mathrm{argmin}_{p \in \mathcal{U}_d^+}\left[-\sum_{s \in \mathcal{M}} \log\left(\frac{p}{\bar{p}_n}\right)\bar{p}_n + \sum_{s \in \mathcal{M}} p\right] - p_0\right\} \\
&= \quad \mathrm{argmin}_{q \in \sqrt{n}(\mathcal{U}_d^+ - p_0)}\left[-\sum_{s \in \mathcal{M}} \log\left(\frac{p_0 + q/\sqrt{n}}{\bar{p}_n}\right)\bar{p}_n + \sum_{s \in \mathcal{M}} \left(p_0 + q/\sqrt{n}\right)\right] \\
&= \quad \mathrm{argmin}_{q \in \sqrt{n}(\mathcal{U}_d^+ - p_0)}\left[-\sum_{s \in \mathcal{M}} \log\left(\frac{p_0 + \frac{q}{\sqrt{n}}}{\bar{p}_n}\right)\bar{p}_n + \sum_{s \in \mathcal{M}} \frac{q}{\sqrt{n}} - \sum_{s \in \mathcal{M}} \frac{\mathbb{W}_n}{\sqrt{n}}\right],
\end{aligned}$$

since $\sum_{s \in \mathcal{M}} p_0$ and $\sum_{s \in \mathcal{M}} \mathbb{W}_n$ are constants on which the minimization does not depend. Now, let

$$\begin{aligned}
\mathbb{M}_n(q) \quad &= \quad -\sum_{s \in \mathcal{M}} \log\left(\frac{p_0 + q/\sqrt{n}}{\bar{p}_n}\right)\bar{p}_n + \frac{1}{\sqrt{n}}\sum_{s \in \mathcal{M}} q - \frac{1}{\sqrt{n}}\sum_{s \in \mathcal{M}} \mathbb{W}_n \\
&= \quad -\sum_{s \in \mathcal{M}} \log\left(1 + \frac{1}{\sqrt{n}}\frac{q - \mathbb{W}_n}{\bar{p}_n}\right)\bar{p}_n + \frac{1}{\sqrt{n}}\sum_{s \in \mathcal{M}} q - \frac{1}{\sqrt{n}}\sum_{s \in \mathcal{M}} \mathbb{W}_n \\
&\approx \quad -\sum_{s \in \mathcal{M}}\left\{\frac{1}{\sqrt{n}}\frac{q - \mathbb{W}_n}{\bar{p}_n} - \frac{1}{2}\left(\frac{1}{\sqrt{n}}\frac{q - \mathbb{W}_n}{\bar{p}_n}\right)^2\right\}\bar{p}_n + \sum_{s \in \mathcal{M}} \frac{q}{\sqrt{n}} - \sum_{s \in \mathcal{M}} \frac{\mathbb{W}_n}{\sqrt{n}} \\
&= \quad \frac{1}{2n}\sum_{s \in \mathcal{M}} \frac{(q - \mathbb{W}_n)^2}{\bar{p}_n} = \widetilde{\mathbb{M}}_n(q)
\end{aligned}$$

where

$$n\widetilde{\mathbb{M}}_n(q) \quad \Rightarrow \quad \frac{1}{2}\sum_{s \in \mathcal{M}} \frac{(q - \mathbb{W})^2}{p_0}.$$

Finally, since p_0 is constant on $\mathcal{M}$, we would therefore like to conclude that

$$\sqrt{n}(\widehat{p}_n - p_0) \quad \Rightarrow \quad \mathrm{argmin}_{q \in \mathcal{U}_{|\mathcal{M}|}} \sum_{s \in \mathcal{M}} (q - \mathbb{W})^2 \quad = \quad \mathrm{uni}[\mathbb{W}_{\mathcal{M}}],$$

where $\mathbb{W}_{\mathcal{M}}$ denotes the vector of random variables $\{\mathbb{W}(s), s \in \mathcal{M}\}$. To do this, we need to check the criteria of the argmax continuous mapping theorem, that is

1. $\sqrt{n}(\widehat{p}_n - p_0)$ is tight ("uniformly tight" in the sense of van der Vaart and Wellner (1996)) since it is equal to $\sqrt{n}(\widehat{p}_n|_{\widehat{\kappa}_n} - p_0)$ for some $\widehat{\kappa}_n \in \mathcal{M}$, and each $\sqrt{n}(\widehat{p}_n|_k - p_0)$ converges, using for example, Marshall's lemma.

2. The requirement that $\mathbb{M}_n(\sqrt{n}(\widehat{p}_n - p_0)) \geq \sup_q \mathbb{M}_n(q)$ is satisfied by definition of the $\widehat{p}_n$.

3. By Lemma B.5, $\sum_{s \in \mathcal{M}} (q - \mathbb{W})^2$ has a unique minimum on $\mathcal{U}_d$, that is, uni$[\mathbb{W}_{\mathcal{M}}]$ has a unique solution. To see this, recall that on $\mathcal{M}$, $\mathbb{W}$ is normally distributed with mean zero and covariance given by $\mathrm{cov}(\mathbb{W}(s_i), \mathbb{W}(s_j)) = \theta\delta_{i,j} - \theta^2$, letting $\theta = p_0(s), s \in \mathcal{M}$. Now, define $\mathbb{V}(s) = (\theta|\mathcal{M}|)^{-1/2}(\mathbb{W}(s) - \sum_{s \in \mathcal{M}} \mathbb{W}(s)/|\mathcal{M}|)$, so that $\mathbb{W} = (\theta d)^{1/2}\mathbb{V} + \sum_{s \in \mathcal{M}} \mathbb{W}(s)/d$, using $d = |\mathcal{M}|$. A quick check shows that $\mathbb{V}$ is still normally distributed with mean zero and $\mathrm{cov}(\mathbb{V}(s_i), \mathbb{V}(s_j)) = d^{-1}\delta_{i,j} - d^{-2}$. Applying Lemma C.5, we have that

$$\mathrm{uni}[\mathbb{W}] \quad = \quad (\theta d)^{1/2}\,\mathrm{uni}[\mathbb{V}] + \sum_{s \in \mathcal{M}} \mathbb{W}(s)/d.$$

   By Lemma B.5, uni$[\mathbb{V}]$ has a unique solution, and therefore, uni$[\mathbb{W}]$ does also.

4. Note lastly that $\sum_{s \in \mathcal{M}} (q - \mathbb{W})^2$ is a.s. continuous in q.

The result follows. □

B.4 Proof of Proposition 5.3

Lemma B.6. *Suppose that* $\sum_{s \in S_0} p_0^{1/2}(s) < \infty$. *Then*

$$\sqrt{n} \sum_{s \notin S_n} p_0(s) = o_p(1).$$

Proof. Let $P_0(A) = \sum_{s \in A} p_0(s)$ and $\mathbb{P}_n(A) = \sum_{s \in A} \overline{p}_n(s)$. By the Borisov-Durst theorem (Dudley, 1999, Theorem 7.9, page 279) the power set of S_0, 2^{S_0}, is a Donsker class for P_0 if and only of $\sum_{s \in S_0} p_0^{1/2}(s) < \infty$. Now, let $\mathbb{G}$ denote the zero-mean Gaussian random field on 2^{S_0} with covariance

$$E\left[\mathbb{G}(A)\mathbb{G}(B)\right] = P_0(A \cap B) - P_0(A)P_0(B).$$

The Borisov-Durst theorem tells us that

$$\sqrt{n}(\mathbb{P}_n - P_0) \quad \Rightarrow \quad \mathbb{G} \quad \text{in } \ell_\infty(2^{S_0}). \tag{B-23}$$

Since S_0 is countable, we have that

$$\sup_{A \in 2^{S_0}} \left| \mathbb{P}_n(A) - P_0(A) \right| = \frac{1}{2}\, \ell_1(\bar{p}_n, p_0) \to 0,$$

almost surely as $n \to \infty$ since the class 2^{S_0} is also Glivenko-Cantelli. Since by definition, $\mathbb{P}_n(S_n^c) = 0$, the latter implies that

$$\lim_{n\to\infty} P_0(S_n^c) = 0 \tag{B-24}$$

almost surely. Furthermore, using the Skorokhod representation the convergence in (B-23) (see e.g. Theorem 1.10.4 of van der Vaart and Wellner, 1996) implies that we can assume that there exists a common probability space on which $\sqrt{n}(\mathbb{P}_n - P_0)$ and $\mathbb{G}$ are defined such that

$$\sup_{A \in 2^{S_0}} \left| \sqrt{n}(\mathbb{P}_n(A) - P_0(A)) - \mathbb{G}(A) \right| \to 0$$

almost surely. This implies that

$$\lim_{n\to\infty} \left(\sqrt{n}(\mathbb{P}_n(S_n^c) - P_0(S_n^c)) - \mathbb{G}_{P_0}(S_n^c) \right) = 0$$

almost surely. However, $\mathbb{G}(S_n^c) \overset{d}{=} Z\sqrt{P_0(S_n^c)(1 - P_0(S_n^c))}$ with $Z \sim \mathcal{N}(0,1)$. Using this along with $\mathbb{P}_n(S_n^c) = 0$ and (B-24) it follows that

$$\lim_{n\to\infty} \sqrt{n} P_0(S_n^c) = 0.$$

We conclude that on the original probability space $\sqrt{n}P_0(S_n^c) \to_d 0$ which is equivalent to

$$\sqrt{n} P_0(S_n^c) = \sqrt{n} \sum_{s \in S_n^c} p_0(s) \overset{p}{\to} 0,$$

because the limit is degenerate. □

Proof of Proposition 5.3. Let us first fix $\kappa \in \mathcal{M}$. From Corollary 4.6 we have that with probability one, there exists a sufficiently large n_0 such that $\widehat{p}_n = \widehat{p}_n|_{\widehat{\kappa}_n}$ with $\widehat{\kappa}_n \in \mathcal{M}$ if $n \geq n_0$. This also implies that $\mathcal{M} \subset S_n$ for all $n \geq n_0$. Consider such an n.

Since $\kappa \in \mathcal{M}$, we know that $p_0 \in \mathcal{U}^1|_{\kappa}(S_0)$. From the characterization of the restricted MLE, we know that $\widehat{F}_{n|k}(s), s \in S_n$ (the associate CDF) is found as the least concave majorant of the graph

$$\{(i, \mathbb{F}_n(z_i)), k \leq i \leq m\}$$

where z_i denotes an ordered enumeration of the elements of $S_n = \{z_1, \ldots, z_m\}$ and $k = k(\kappa, S_n)$ is such that $z_k = s_\kappa$ (recalling that $\mathcal{M} \subset S_n$). Next, define the function

$$\overline{F}_0(z_i) \;=\; \sum_{j \leq i} p_0(z_j).$$

This depends of course on the observed S_n. Note that by definition this function is concave on $k \le i \le m$ and convex on $1 \le i \le k-1$. Now, the usual proof of Marshall's lemma applies. That is, let $a = \sup_{i\ge k} |\mathbb{F}_n(z_i) - \overline{F}_0(z_i)|$. Then for all $i \ge k$, we have (1)

$$\widehat{F}_{n|k}(z_i) - \overline{F}_0(z_i) \;\ge\; \mathbb{F}_n(z_i) - \overline{F}_0(z_i) \;\ge\; -a.$$

On the other hand $\overline{F}_0(z_i) + a$ is a concave majorant of $\mathbb{F}_n(z_i)$ on $i \ge k$, and hence (2) $\overline{F}_0(z_i) + a \ge \widehat{F}_n(z_i)$. Combining the results of (1) and (2) gives that $\sup_{i\ge\kappa} |\widehat{F}_n(z_i) - \overline{F}_0(z_i)| \le \sup_{i\ge\kappa} |\mathbb{F}_n(z_i) - \overline{F}_0(z_i)|$. Repeating the argument on greatest concave minorants, yields a similar result for $i \le \kappa - 1$, which combined gives

$$\sup_{z\in S_n} |\widehat{F}_{n|k}(z) - \overline{F}_0(z)| \;\le\; \sup_{z\in S_n} |\mathbb{F}_n(z) - \overline{F}_0(z)|.$$

This result holds for any choice of $\kappa \in \mathcal{M}$. Next, from Corollary 4.6 we have that with probability one, there exists a sufficiently large n_0 such that $\widehat{p}_n = \widehat{p}_n|_{\widehat{\kappa}_n}$ with $\widehat{\kappa}_n \in \mathcal{M}$. Let $\widehat{F}_n$ denote the CDF associated with $\widehat{p}_n$. We then have that

$$\begin{aligned} \sup_{z\in S_n} |\widehat{F}_n(z) - \overline{F}_0(z)| &\le \sup_{\kappa\in\mathcal{M}} \sup_{z\in S_n} |\widehat{F}_{n|k}(z) - \overline{F}_0(z)| \\ &\le \sup_{z\in S_n} |\mathbb{F}_n(z) - \overline{F}_0(z)|. \end{aligned}$$

Next, it follows that

$$\begin{aligned} \sup_{s\in S_n} |\widehat{F}_n(s) - \overline{F}_0(s)| &\le \sup_{s\in S_n} |\mathbb{F}_n(s) - F_0(s)| + \sup_{s\in S_n} |F_0(s_i) - \overline{F}_0(s_i)| \\ &\le \sup_{s\in S_0} |\mathbb{F}_n(s) - F_0(s)| + \sum_{s\in S_n^c} p_0(s). \end{aligned}$$

On the other hand,

$$\begin{aligned} |\widehat{F}_n(s) - \overline{F}_0(s)| &= |\widehat{F}_n(s) - F_0(s) + F_0(s) - \overline{F}_0(s)| \\ &\ge |\widehat{F}_n(s) - F_0(s)| - |F_0(s) - \overline{F}_0(s)| \end{aligned}$$

This yields

$$\sup_{s\in S_0} |\widehat{F}_n(s) - F_0(s)| \;\le\; \sup_{s\in S_0} |\mathbb{F}_n(s) - F_0(s)| + 2 \sum_{s\in S_n^c} p_0(s).$$

The full result is obtained by applying Lemma B.6. □

B.5 Proofs for Section 6

Proof of Proposition 6.1. Using our assumption of finite support, the result follows immediately from Theorems 4.4 and 5.1 via Slutsky's theorem and the continuity of norms on $\mathbb{R}^d$. Next, recall the definition of φ in (5.15). We have that

$$\frac{\varphi(\mathbb{W})}{p_0^\beta} = \varphi\left(\frac{\mathbb{W}}{p_0^\beta}\right)$$

since p_0 is constant on the intervals $\mathcal{I}_j, \mathcal{M}, \mathcal{D}_j$. The final inequality now follows as in Proposition 5.2. □

Proposition B.7. *Let $\mathbb{W}$ denote a mean zero Gaussian process defined on S_0 such that $cov(\mathbb{W}(s_i), \mathbb{W}(s_j)) = p_0(s_i)\delta_{i,j} - p_0(s_i)p_0(s_j),\ s_i \in S_0$. Let $\widetilde{\mathbb{W}}_n$ denote a mean zero Gaussian process defined on $supp(\widehat{p}_n)$ such that $cov(\widetilde{\mathbb{W}}_n(s_i), \widetilde{\mathbb{W}}_n(s_j)) = \widehat{p}_n(s_i)\delta_{i,j} - \widehat{p}_n(s_i)\widehat{p}_n(s_j),\ s_i \in supp(\widehat{p}_n)$. Let $q_{0,\alpha}$ and $\widetilde{q}_{0,\alpha}$ denote the quantiles such that*

$$P(\|\mathbb{W}\|_\infty > q_{0,\alpha}) = \alpha, \qquad P(\|\widetilde{\mathbb{W}}_n\|_\infty > \widetilde{q}_{0,\alpha}) = \alpha,$$

respectively. Then $\widetilde{q}_{0,\alpha} \to q_{0,\alpha}$ almost surely.

Proof. First, let p_n denote *any fixed* pmf such that p_n converges to p_0 and has the same properties as $\widehat{p}_n$:

(a) p_n converges pointwise to p_0, and

(b) $\lim_{m\to\infty} \lim_n \sum_{|s_i|>m} p_n(s_i) = 0$.

Suppose also that $\widetilde{\mathbb{W}}_n$ is defined as above, except that p_n replaces $\widehat{p}_n$ in the definition (in essence, we remove the randomness associated with this choice). Then one can easily show that $\widetilde{\mathbb{W}}_n$ converges weakly to $\mathbb{W}$ in ℓ_2. This follows from (a) convergence of finite dimensional distributions, which is immediate from convergence of p_n to p_0, and (b) tightness in ℓ_2. To prove tightness, we refer again to Jankowski and Wellner (2009, Lemma 6.2). Note that we have that

1. $E[\|\widetilde{\mathbb{W}}_n\|_2^2] \le 1$ for all n
2. For sufficiently large n, we have that

$$\sum_{|s_i|>m} E[\widetilde{\mathbb{W}}_n^2(s_i)] \le \sum_{|s_i|>m} p_n(s_i),$$

which shows that $\widetilde{\mathbb{W}}_n$ is tight in ℓ_2. The required weak convergence follows. Now, since the ℓ_∞ is continuous in ℓ_2, convergence of the quantiles follows.

Thus, we obtain convergence of the quantiles (as numbers), based on conditions (a) and (b) of p_n. We will now show that these conditions hold almost surely, establishing the full result. Condition (a) follows immediately from Theorem 4.4. To see also that Condition (b) holds, note that from

Propositions C.2 and B.4, there exists a sufficiently large n such that with probability one

$$\sum_{|s_i|>m} \widehat{p}_n(s_i) \quad = \quad \widehat{F}_n(-m) + 1 - \widehat{F}_n(m) \quad \leq \quad \mathbb{F}_n(-m) + 1 - \mathbb{F}_n(m)$$

for $m \notin \mathcal{M}$. That $\lim_m \lim_n(\mathbb{F}_n(-m) + 1 - \mathbb{F}_n(m)) = 0$ almost surely follows from the properties of the empirical CDF and CDFs in general. □

C Additional Technical Results

C.1 Useful bounds

Lemma C.1. *Any $p \in \mathcal{U}^1|_\kappa(S_0)$ satisfies*

$$p(s_j) \leq \min\left\{1, |j-\kappa|^{-1}\right\}.$$

Proof. We have that

$$1 \quad \geq \quad \sum_{i=\kappa}^{j} p(s_i) \quad \geq \quad \sum_{i=\kappa}^{j} p(s_j) \quad = \quad (j-\kappa+1)p(s_j) \quad \geq \quad (j-\kappa)p(s_j).$$

Similarly, we have

$$1 \quad \geq \quad \sum_{i=j}^{\kappa-1} p(s_i) \quad \geq \quad \sum_{i=j}^{\kappa-1} p(s_j) \quad = \quad (\kappa-j)p(s_j).$$

Together, these yield the first inequality. □

Proposition C.2. *The restricted MLE $\widehat{p}_n|_\kappa$ satisfies the inequalities*

$$\begin{array}{lll} \widehat{F}_n|_\kappa(z) & \geq & \mathbb{F}_n(z) \quad z \geq s_\kappa, \\ \widehat{F}_n|_\kappa(z) & \leq & \mathbb{F}_n(z) \quad z \leq s_{\kappa-1}. \end{array}$$

Proof. Follows immediately from the GCM/LCM characterization of $\widehat{p}_n|_\kappa$. □

C.2 Proof of Proposition 2.1

Suppose that there exists q such that p satisfies (2.4). It is clear that p is a pmf. We now verify that p is unimodal with mode either at $s_{\kappa-1}$ or s_κ. Let $(\Delta p)(j) = p(s_{j+1}) - p(s_j)$. We calculate

$$(\Delta p)(j) \quad = \quad \begin{cases} -\frac{q(s_{j-\kappa})}{j-\kappa+1} & \leq \; 0 \quad j \geq \kappa, \\ \frac{q(s_{j+1-\kappa})}{|j+1-\kappa|} & \geq 0 \quad j \leq \kappa-2. \end{cases}$$

Therefore, p is non-decreasing on $\{s_j : j \geq \kappa\}$ and non-increasing on $\{s_j : j \leq \kappa - 1\}$. For $j = \kappa - 1$, we calculate

$$p(s_\kappa) - p(s_{\kappa-1}) = \sum_{i=\kappa}^{\infty} \frac{q(s_i)}{i - \kappa + 1} - \sum_{-\infty}^{i=\kappa-1} \frac{q(s_i)}{|i - \kappa|}$$

which could be either ≥ 0 or < 0. This shows that p is unimodal with mode either at $s_{\kappa-1}$ or s_κ.

Conversely, if p is a pmf which is unimodal with mode either at $s_{\kappa-1}$ or s_κ. Let q be defined as

$$q(s_i) \quad = \quad \begin{cases} -(i - \kappa + 1)(\Delta p)(i) & i \geq \kappa, \\ |i - \kappa|(\Delta p)(i - 1) & i \leq \kappa - 1. \end{cases}$$

By the property of p, $q \geq 0$. Furthermore, using Fubini's theorem and the fact that p is a pmf, we have that

$$\begin{aligned} \sum_j q(s_j) &= -\sum_{j \geq \kappa} (j - \kappa + 1)(\Delta p)(j) + \sum_{j \leq \kappa - 1} (\kappa - j)(\Delta p)(j - 1) \\ &= -\sum_{i=0}^{\infty} \sum_{j=i+\kappa}^{\infty} (p(s_{j+1}) - p(s_j)) + \sum_{i=0}^{\infty} \sum_{j=-\infty}^{\kappa-1-i} (p(s_j) - p(s_{j-1})) \\ &= \sum_{i=0}^{\infty} p(s_{i+\kappa}) + \sum_{i=0}^{\infty} p(s_{\kappa-1-i}) \\ &= \sum_{i \geq \kappa} p(s_i) + \sum_{i \leq \kappa - 1} p(s_i) = 1 \end{aligned}$$

and hence q is a pmf. Finally, q satisfies

$$\begin{aligned} \sum_i (|i| + 1)^{-1} q(s_{i+\kappa}) &= \sum_{i \geq \kappa} \frac{q(s_i)}{i - \kappa + 1} + \sum_{i \leq \kappa - 1} \frac{q(s_i)}{\kappa - i} \\ &= -\sum_{i \geq \kappa} (\Delta p)(i) + \sum_{i \leq \kappa - 1} (\Delta p)(i - 1) \\ &= p(s_\kappa) + p(s_{\kappa-1}) < \infty \end{aligned}$$

which completes the proof. □

C.3 Properties of the anti, iso, and uni operators

There is a well-known equivalence between the monotonic projection in the sense of least squares and likelihood maximization (e.g. the maximum likelihood and least squares estimators are the same for a decreasing density). As such equivalences are not always readily available in a standard reference on isotonic estimation, for completeness, we state this relationship explicitly in the following lemma. Let $\mathcal{I}_d^+ = \mathcal{I}_d \cap \{u \in \mathbb{R}^d : u_j > 0\}$ and $\mathcal{D}_d^+ = \mathcal{D}_d \cap \{u \in \mathbb{R}^d : u_j > 0\}$.

Lemma C.3. *Suppose that $v \in \mathbb{R}^d$ such that $v_j > 0$ for $j = 1, \dots, d$. Then*

$$\begin{aligned} \text{iso}[v] &= \text{argmax}_{u \in \mathcal{I}_d^+} \Big\{ \sum_{j=1}^d v_j \log(u_j) - \sum_{j=1}^d u_j \Big\}, \\ \text{anti}[v] &= \text{argmax}_{u \in \mathcal{D}_d^+} \Big\{ \sum_{j=1}^d v_j \log(u_j) - \sum_{j=1}^d u_j \Big\}. \end{aligned}$$

Proof. It is known that

$$\text{argmin}_{u \in \mathcal{I}_d} \sum_{j=1}^d (v_j - u_j)^2$$

is equal to the right slope of the GCM of the cumulative sum diagram $\big\{(0,0), (j, \sum_{i=1}^j v_j), j = 1, \dots, d\big\}$. Note that implies in particular that these slopes have to be positive if $v_j > 0$ for all $j \in \{1, \dots, d\}$, and hence

$$\text{argmin}_{u \in \mathcal{I}_d} \sum_{j=1}^d (v_j - u_j)^2 = \text{argmin}_{u \in \mathcal{I}_d^+} \sum_{j=1}^d (v_j - u_j)^2.$$

Now maximizing the criterion $L(u) = \sum_{j=1}^d v_j \log(u_j) - \sum_{j=1}^d u_j$ on $\mathcal{I}_d^+$ admits a unique solution. Let $\{u^s\}_{s \in \mathbb{N}}$ be a maximizing sequence of L. Suppose that there exists $j \in \{1, \dots, d\}$ such that

$$\lim_{s \to \infty} u_j^s = 0 \quad \text{or} \quad \lim_{s \to \infty} u_j^s = \infty.$$

Then, in this case we would have $\lim_{s \to \infty} L(u^s) = -\infty$ contradicting the fact that $\{u^s\}_{s \in \mathbb{N}}$ is a maximizing sequence since it must satisfy $\lim_{s \to \infty} L(u^s) \geq L(v) = \sum_{j=1}^d v_j \log(v_j) - \sum_{j=1}^d v_j > -\infty$. Hence, there exists $K_2 > K_1 > 0$ such that $K_1 \leq u_j \leq K_2$ for $j = 1, \dots, d$. It follows that the maximization is performed on a compact set and existence of the maximum is now guaranteed by continuity of L. Uniqueness follows from strict concavity of L. We denote this unique solution by $\widehat{v}$. Let $j \in \{1, \dots, d\}$. For $\epsilon \in \mathbb{R}$, let

$$\widehat{v}_i^\epsilon = \widehat{v}_i + \epsilon \mathbb{I}_{1 \leq i \leq j}, \; 1 \leq i \leq d.$$

Then, for $\epsilon > 0$ small enough, we have $\widehat{v}^\epsilon \in \mathcal{I}_d^+$, and $L(\widehat{v}^\epsilon) \leq L(\widehat{v})$. Therefore

$$\lim_{\epsilon \searrow 0} \epsilon^{-1} \left(L(\widehat{v}^\epsilon) - L(\widehat{v}) \right) \leq 0.$$

When j is a knot point, that is, $\widehat{v}_{j+1} > \widehat{v}_j$ then it is easy to see that

$$\lim_{\epsilon \to 0} \epsilon^{-1} \left(L(\widehat{v}^\epsilon) - L(\widehat{v}) \right) = 0.$$

This yields

$$\sum_{i=1}^j \frac{v_i}{\widehat{v}_i} \begin{cases} \leq j, & \text{for all } j \in \{1, \dots, d\} \\ = j, & \text{if } j \text{ is a knot point.} \end{cases} \tag{C-25}$$

Let $B_1, \ldots, B_r$ denote a partition of $\{1, \ldots, d\}$ such that $\forall l \in B_i$, $u_l = c_i$ some positive constant, for $i = 1, \ldots, r$. Let $i_1, i_2, \ldots, i_r$ denote the largest integers of $B_1, \ldots, B_d$ respectively. Note that $i_r = d$. Then, if follows from (C-25) that

$$\sum_{i=1}^{j} v_i \begin{cases} \leq j\widehat{v}_1 = \sum_{i=1}^{j} \widehat{v}_i, \text{ for all } j \in B_1 = \{1, \ldots, i_1\} \\ = j\widehat{v}_1 = \sum_{i=1}^{j} \widehat{v}_i, \text{for } j = i_1. \end{cases}$$

The same reasoning can be applied for the other sets $B_i, 2 \leq i \leq r$ to conclude that

$$\sum_{i=1}^{j} \widehat{v}_i \begin{cases} \geq \sum_{i=1}^{j} v_i, \text{for all } j \in \{1, \ldots, d\} \\ = \sum_{i=1}^{j} v_i, \text{if } j \text{ is a jump point.} \end{cases} \tag{C-26}$$

Hence, the solution $\widehat{v}$ is given by the slope of the LCM of the cumulative sum of v. The same reasoning can be applied to the projection on $\mathcal{D}_d^+$, proving the result. □

In the following, we state a result which shows that isotonic/antitonic projections can be transformed into "localized"projections between the knots of the "global" isotonic/antitonic solution. Recall that if $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, then $v_{s:t} = (v_s, \ldots, v_t)$ for $1 \leq s \leq t \leq d$.

Lemma C.4. *Let $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ such that $v_j > 0, \ j = 1, \ldots, d$. Also let $\widehat{v} = \text{iso}[v]$ and $1 \leq s_1 < \ldots < s_r \leq d$ the locations of the knot points of $\widehat{v}$, that is*

$$\widehat{v}_1 = \ldots = \widehat{v}_{s_1} < \widehat{v}_{s_1+1} = \ldots \widehat{v}_{s_2} < \ldots < \widehat{v}_{s_r+1} = \ldots = \widehat{v}_d.$$

Then, for $1 \leq j < k \leq r$

$$\widehat{v}_{(s_j+1):s_k} = \text{iso}[v_{(s_j+1):s_k}].$$

Proof. The proof follows immediately from the fact that $\widehat{v}_{(s_j+1):s_k}$ is characterized by the same Fenchel conditions as $\text{iso}(v_{(s_j+1):s_k})$. Indeed, we know from the characterization of $\widehat{v} = \text{iso}(v)$ that

$$\sum_{i=1}^{t} \widehat{v}_i \begin{cases} \geq \sum_{i=1}^{t} v_i, \text{for all } t \in \{1, \ldots, d\} \\ = \sum_{i=1}^{t} v_i, \text{if } t \text{ is a jump point} \end{cases}$$

and therefore

$$\sum_{i=s_j}^{t} \widehat{v}_i \begin{cases} \geq \sum_{i=s_j}^{t} v_i, \text{for all } t \in \{s_j + 1, \ldots, s_k\} \\ = \sum_{i=s_j}^{t} v_i, \text{if } t \text{ is a jump point} \end{cases}$$

which give exactly the characterization of the isotonic projection of the subvector $v_{(s_j+1):s_k}$.

□

Lemma C.5. *Suppose that* $v \in \mathbb{R}^d$ *and let* $p \in \mathcal{I}_d, q \in \mathcal{D}_d$*. Also, let* $a > 0, b \in \mathbb{R}$ *denote two fixed constants. Then the following (in)equalities hold*

$$\begin{array}{rclrcl} \| \operatorname{iso}[v] - p\|_2^2 & \leq & \|v - p\|_2^2, & \| \operatorname{anti}[v] - q\|_2^2 & \leq & \|v - q\|_2^2 \\ \operatorname{anti}[av + b] & = & a \operatorname{anti}[v] + b, & \operatorname{iso}[av + b] & = & a \operatorname{iso}[v] + b, \\ \operatorname{uni}[av + b] & = & a \operatorname{uni}[v] + b. & & & \end{array}$$

Proof. The first two inequalities appear in Robertson et al. (1988, Theorem 1.6.1); cf. Jankowski and Wellner (2009, Lemma 6.1). The three equalities are all proved in a similar manner. For example,

$$\begin{array}{rcl} \operatorname{anti}[v + b] & = & \operatorname{argmin}_{u \in \mathcal{D}_d} \|u - (v + b)\|_2^2 \\ & = & \operatorname{argmin}_{u \in \mathcal{D}_d} \|(u - b) - v\|_2^2 \\ & = & \operatorname{argmin}_{u + b \ \in \mathcal{D}_d} \|u - v\|_2^2 \\ & = & \operatorname{argmin}_{u \in \mathcal{D}_d} \|u - v\|_2^2 + b \quad = \quad \operatorname{anti}[v] + b. \end{array}$$

□

Continuity of the operators anti and iso follows immediately from Jankowski and Wellner (2009, Lemma 6.1).

Proposition C.6. *Suppose that* $v_n \in \mathbb{R}^d$ *and that* $\lim_{n \to \infty} v_n = v$*. Then*

$$\lim_{n \to \infty} \operatorname{iso}[v_n] = \operatorname{iso}[v], \quad \textit{and} \quad \lim_{n \to \infty} \operatorname{anti}[v_n] = \operatorname{anti}[v].$$