Bernoulli **23**(3), 2017, 1449–1480
DOI: 10.3150/15-BEJ754

On asymptotics of the discrete convex LSE of a p.m.f.

FADOUA BALABDAOUI[1], CÉCILE DUROT[2] and FRANÇOIS KOLADJO[3,4]

[1]*Université Paris-Dauphine, PSL Research University, CNRS, CEREMADE, 75016 Paris, France. E-mail: fadoua@ceremade.dauphine.fr*
[2]*Modal'x, Université Paris Nanterre, 92001 Nanterre, France. E-mail: cecile.durot@gmail.com*
[3]*Equipe Probabilité, Statistique et Modélisation, UMR CNRS 8628, Université Paris-Sud, 91405 Orsay Cedex, France. E-mail: francois.koladjo@gmail.com*
[4]*CIPMA-Chaire UNESCO, FAST, UAC, 072BP50 Cotonou, Bénin.*

In this article, we derive the weak limiting distribution of the least squares estimator (LSE) of a convex probability mass function (p.m.f.) with a finite support. We show that it can be defined via a certain convex projection of a Gaussian vector. Furthermore, samples of any given size from this limit distribution can be generated using an efficient Dykstra-like algorithm.

Keywords: convex; least squares; nonparametric estimation; p.m.f.; shape-constraints

1. Introduction

Non-parametric estimation under a shape constraint of a density on a given sub-interval of $\mathbb{R}$, has attracted considerable attention over the past decades. Typical shape constraints are monotonicity, convexity, log-concavity. Typical estimators are the maximum likelihood estimator (MLE) and the least-squares estimator (LSE). Both of them are obtained by minimization of a given criterion over the set of all densities that satisfy the considered shape constraint. Even if the MLE and LSE uniquely exist, no closed form is available for these estimators so a key step is to provide a precise characterization of the estimators as well as an algorithm for practical implementation. Grenander [12] first gives such a characterization for the MLE of a monotone density, and the pointwise weak convergence of the MLE is derived in [20]. The estimator can easily be implemented using the Pool Adjacent Violators Algorithm as described in [4]. Both characterization and pointwise weak convergence of a convex density estimator on the half-real line are investigated in [13], and practical implementation is discussed in [14]. The MLE of a log-concave density is characterized in [5] while its pointwise weak convergence is studied in [3]. Algorithmic aspects are treated in [6]. In the aforementioned continuous case of estimating a density under a shape constraint over a given sub-interval of $\mathbb{R}$, the limit behavior of global distances between the estimator and the true density has been investigated. We refer to [17] and [10] for the limit distribution of the L_p-distance and the supremum-distance respectively, in the case of a monotone density. The rate of uniform convergence of the log-concave MLE on compacts is given in [5].

More recently, attention has been given to estimation of a discrete probability mass function (p.m.f.) under a shape constraint. Similar to the continuous case, no closed form is available for

1350-7265

shape constrained estimators of a p.m.f. so one needs a precise characterization. Such characterizations are given in [2,16] and [8] for the monotone, log-concave and convex cases respectively. In the convex case, [1] show that the same Marshall Lemma proved in the continuous case by [7] continues to hold in the discrete case, that is, we have $\|\widehat{F}_n - F\|_\infty \leq 2\|\mathbb{F}_n - F\|_\infty$ for any distribution function F on $\mathbb{N}$ with a convex p.m.f., and $\widehat{F}_n$ and $\mathbb{F}_n$ are the cumulative distribution function of the LSE and the empirical distribution function, respectively. In contrast to the continuous case, the natural way to investigate the global limit behavior of the estimator is to compute the limit distribution of the whole process $\widehat{p}_n - p_0$, where $\widehat{p}_n$ is the considered estimator and p_0 is the true p.m.f. This approach was first considered by [16] in the case of a monotone p.m.f. on $\mathbb{N}$, and by [2] in the case of a log-concave p.m.f. The discrete case totally differs from the continuous case. In particular, the rate of convergence is typically $\sqrt{n}$ (where n denotes the sample size) in the discrete case whereas it is of smaller order in the continuous case. Characterization and rate of convergence of the LSE of a convex p.m.f. are given in [8] together with an algorithm, but the limit distribution of the LSE remains unknown. One of the aims of this paper is to fill this gap.

To be more precise, let $X_1, \ldots, X_n$ be i.i.d. from an unknown discrete p.m.f. p_0 whose support takes the form $\mathbb{N} \cap [\kappa, \infty) = \{\kappa, \kappa + 1, \ldots\}$ or $\mathbb{N} \cap [\kappa, S] = \{\kappa, \kappa + 1, \ldots, S\}$ for some integers $S > \kappa \geq 0$. Here, κ is assumed to be known whereas S is unknown. Assuming that p_0 is convex on $\mathbb{N} \cap [\kappa, \infty)$, we are interested in the limiting behavior of the LSE of p_0. The case $\kappa = 1$ is of a particular interest in [9], where the problem of estimating the total number N of species in a given area is investigated under the convexity constraint. Note that we focus here on the LSE. Studying the limit behavior of the MLE is out of the scope of the article: it would require specific arguments since the MLE may differ from the LSE in our setup, see Section 2.2 below.

The limiting distribution of the LSE is described as a piecewise convex projection of a Gaussian process, where the pieces are connected to the points of strict convexity of p_0. The Gaussian process involved in the limiting distribution depends on p_0 as well. Hence, we provide an estimator of the limiting distribution that involves consistent estimators of the points of strict convexity of p_0. We provide an algorithm for simulating the limiting distribution of the LSE as well as the approximating distribution. This amounts to simulating (many times) a Gaussian process and its piecewise convex projection, which is obtained by minimizing the least-squares criterion over the intersection of closed convex cones. Our algorithm combines two previous algorithms. The first one is implemented in the function `conreg` of the package `cobs` for R; see [19] for a full description. It is used to minimize the least squares criterion over the closed convex cone of discrete convex functions on a given interval. Then, the iterative algorithm by [11] is used to minimize the criterion over the intersection of closed convex cones. We use our algorithm to illustrate our main results via a simulation study.

The paper is organized as follows. In Section 2, we recall the characterization of the LSE obtained in [8] and we derive the $\sqrt{n}$-rate of convergence. We show that the MLE and the LSE may differ, and that the MLE may be non unique. In addition, we prove that any knot (that is, a point of strict convexity) of the true p.m.f. is also (almost surely, for large enough n) a knot of the LSE. This allows us to characterize the support of the LSE in the case when the true p.m.f. has a finite support. Section 3 is devoted to the weak convergence of the LSE. The limit distribution is computed in the general case and we investigate how the limit distribution simplifies in some specific cases, such as p.m.f. having consecutive knots. Simulations are reported in Section 4. We

first investigate on a few examples whether the knots of the estimator include all true knots when the sample size is finite. Then, we illustrate the convergence of the distribution of the estimator to the limit distribution. All proofs are postponed to Section 5.

2. Basic properties of the convex LSE

Let $X_1, \ldots, X_n$ be i.i.d. from a p.m.f. p_0 with support included in $\mathbb{N}$. Denoting by κ the left endpoint of the support, we assume that p_0 is convex on $\mathbb{N} \cap [\kappa, \infty)$. This means that the support of p_0 takes either the form $\mathbb{N} \cap [\kappa, \infty) = \{\kappa, \kappa + 1, \ldots\}$ or $\mathbb{N} \cap [\kappa, S] = \{\kappa, \kappa + 1, \ldots, S\}$ for some integers $S > \kappa \geq 0$, and that $\Delta p_0(k) \geq 0$ for all integers $k > \kappa$, where for a sequence $p = \{p(k), k \in \mathbb{N}\}$,

$$\Delta p(k) = p(k+1) - 2p(k) + p(k-1), \qquad k \in \mathbb{N} \setminus \{0\} \tag{2.1}$$

denotes the corresponding discrete Laplacian. Here, κ is assumed to be known whereas S is unknown. The assumption $S > \kappa$ is made in order to avoid the uninteresting situation of having to deal with a Dirac distribution. Since convexity is preserved under translation, we can assume without loss of generality that $\kappa = 0$: in case $\kappa > 0$, the characterization as well as the asymptotic results for the LSE of the true convex p.m.f. can be easily deduced from the ones established below using the simple fact that the support p.m.f. of $X_i - \kappa$ admits 0 as its left endpoint. Thus, in the sequel we restrict our attention to the case of a convex p.m.f. p_0 on $\mathbb{N}$ with an unknown support.

Based on the sample $X_1, \ldots, X_n$, we consider the empirical p.m.f. p_n, that is

$$p_n(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i = j\}}, \qquad j \in \mathbb{N}. \tag{2.2}$$

2.1. Characterization of the convex LSE

We are mainly interested in the asymptotics of $\widehat{p}_n$, the LSE of p_0 defined as the unique minimizer of the criterion

$$\Phi_n(p) = \frac{1}{2} \sum_{j \in \mathbb{N}} \big(p_n(j) - p(j)\big)^2$$

over $p \in \mathcal{C}$, the set of all convex sequences p on $\mathbb{N}$ with a finite ℓ_2-norm, that is, the set of all sequences $p = \{p(k), k \in \mathbb{N}\}$ satisfying

$$\sum_{k=0}^{\infty} \big|p(k)\big|^2 < \infty \quad \text{and} \quad \Delta p(k) \geq 0 \qquad \text{for all integers } k \geq 1. \tag{2.3}$$

Note that any $p \in \mathcal{C}$ is non-negative and non-increasing. Existence and uniqueness of $\widehat{p}_n$ follows from the Hilbert projection Theorem, see [8], Section 2.1.

It follows from Theorem 1 of [8] that $\widehat{p}_n$ also minimizes Φ_n over the set of p.m.f.'s in $\mathcal{C}$. In particular, $\widehat{p}_n$ is a proper p.m.f. This fact is rather convenient because it means that in order to compute the estimator, we can minimize the criterion Φ_n over $\mathcal{C}$ rather than over the more constrained set of p.m.f.'s in $\mathcal{C}$. This also allows us to use simpler algorithms and has the advantage of giving more flexibility when deriving the characterizing Fenchel conditions for $\widehat{p}_n$. As it is the case in many shape constrained problems, such characterization proves to be crucial in understanding the limiting behavior of the relevant estimator; see, for example, [13,16] and [2]. Thus, for the sake of completeness, we give in Proposition 2.1 below the Fenchel characterization proved in [8], Lemma 2. For an arbitrary sequence $p = \{p(k), k \in \mathbb{N}\}$, we denote

$$F_p(k) = \sum_{j=0}^{k} p(j) \tag{2.4}$$

for all $k \in \mathbb{N}$ with $F_p(-1) = 0$, and we define

$$H_p(z) = \sum_{k=0}^{z-1} F_p(k) \tag{2.5}$$

for all $z \in \mathbb{N}$ with the convention $H_p(0) = 0$. Moreover, a point $k \geq 1$ in the support of a convex $p \in \mathcal{C}$ is called a knot of p if $\Delta p(k) > 0$.

Proposition 2.1. *The convex p.m.f. $\widehat{p}_n$ is the LSE if and only if*

$$H_{\widehat{p}_n}(z) \begin{cases} \geq H_{p_n}(z), & \text{for all } z \in \mathbb{N}, \\ = H_{p_n}(z), & \text{if } z \text{ is a knot of } \widehat{p}_n. \end{cases} \tag{2.6}$$

Note that a typographical error occurring in [8], Lemma 2, is now corrected. More precisely, if $p \in \mathcal{C}$ satisfies $H_p(z) \geq H_{p_n}(z)$ with equality at any knot of p (instead of $\widehat{p}_n$ as stated in [8]) then $p = \widehat{p}_n$.

Some remarks are in order. The characterization above can be seen as the discrete version of the one given by [13] for the LSE of a convex density with respect to Lebesgue measure. However, some of the consequences implied by the characterization of the continuous LSE do not hold true in our discrete case, due to the lack of the notion of differentiability in the discrete case. For instance, if $F_{\widehat{p}_n}$ and F_{p_n} denote the continuous versions of the quantities defined above then $F_{\widehat{p}_n}(s) = F_{p_n}(s)$ at any knot s of the estimator; see Corollary 2.1 of [13]. This equality cannot be expected to hold true for the discrete convex LSE in the general case. In fact, by definition of H_{p_n} and $H_{\widehat{p}_n}$ in the discrete case,

$$F_{p_n}(z) = H_{p_n}(z+1) - H_{p_n}(z) \quad \text{and} \quad F_{\widehat{p}_n}(z) = H_{\widehat{p}_n}(z+1) - H_{\widehat{p}_n}(z)$$

for all $z \in \mathbb{N}$, so it follows from (2.6) that the equality is replaced instead by the two inequalities $F_{\widehat{p}_n}(s) \geq F_{p_n}(s)$ and $F_{\widehat{p}_n}(s-1) \leq F_{p_n}(s-1)$. The equality $F_{\widehat{p}_n}(s) = F_{p_n}(s)$ can only hold if, in addition to the equality $H_{\widehat{p}_n}(s) = H_{p_n}(s)$, one also has $H_{\widehat{p}_n}(s+1) = H_{p_n}(s+1)$. This happens for instance in situations where $\widehat{p}_n$ has two consecutive knots at s and $s+1$.

2.2. The convex MLE compared to the LSE

Recall that the MLE of p_0 is defined as the maximizer (if it exists) of the criterion

$$\ell_n(p) = \sum_{i \geq 0} p_n(i) \log\bigl[p(i)\bigr]$$

over the set of all convex p.m.f.'s p on $\mathbb{N}$, where p_n denotes the empirical p.m.f. The following proposition shows that the LSE of a convex p.m.f. may differ from the MLE. Moreover, it proves that the MLE may be non unique in this discrete setting. To see this, we consider a sample of only one observation X_1 from a convex p.m.f. p_0. We describe the LSE and the MLE in terms of the triangular distributions defined as follows. Given $j \in \mathbb{N} \setminus \{0\}$, consider the triangular p.m.f. with support on $\{0, \ldots, j-1\}$ given by

$$T_j(i) = \frac{2(j-i)_+}{j(j+1)}, \tag{2.7}$$

where as usual, $x_+ = \max\{x, 0\}$ for all real numbers x.

Proposition 2.2. *The convex LSE of p_0 based on the single observation $X_1 > 0$ is uniquely defined by T_{3X_1+1} whereas the MLE exists but is non unique: the* log*-likelihood ℓ_1 is maximized at T_{2X_1}, T_{2X_1+1} and at every mixture of those two triangular distributions.*

These results differ from those obtained by [13] in the continuous case. In that case, the MLE is uniquely defined.

2.3. Tightness of the convex LSE

Next, we consider almost sure consistency of $\widehat{p}_n$ in all distances ℓ_r. Here as usual, $\|p\|_r$ denotes the ℓ_r-norm of a sequence $p = \{p(k), k \in \mathbb{N}\}$:

$$\|p\|_r = \begin{cases} \left(\sum_{k=0}^{\infty} \bigl|p(k)\bigr|^r \right)^{1/r}, & \text{if } r \in \mathbb{N} \setminus \{0\}, \\ \sup_{k \in \mathbb{N}} \bigl|p(k)\bigr|, & \text{if } r = \infty. \end{cases}$$

Proposition 2.3. *For any integer $r \in [2, \infty]$, with probability one we have that*

$$\lim_{n \to \infty} \|\widehat{p}_n - p_0\|_r = 0.$$

The following proposition is an easy consequence of Proposition 2.3.

Proposition 2.4. *If $s > 0$ is a knot of p_0, then with probability one, there exists n_0 such that for all $n \geq n_0$, s is a knot of $\widehat{p}_n$.*

We finish this section by recalling boundedness in probability of $\sqrt{n}(\widehat{p}_n - p_0)$ and the implied boundedness for the associated "integral" processes. Note that boundedness in probability of $\sqrt{n}(\widehat{p}_n - p_0)$ is much weaker than tightness. However, these properties are equivalent in the case where $\sqrt{n}(\widehat{p}_n - p_0)$ is identically equal to zero after a certain range. We will use this equivalence later on under the assumption that the true convex p.m.f. has a finite support.

Theorem 2.5. *If $\widehat{p}_n$ is the convex LSE of the true p.m.f. p_0, then*

$$\sqrt{n}\|\widehat{p}_n - p_0\|_\infty = O_p(1). \tag{2.8}$$

Furthermore, if p_0 has a finite support, then

$$\sqrt{n}\|F_{\widehat{p}_n} - F_{p_0}\|_\infty = O_p(1) \quad \text{and} \quad \sqrt{n}\|H_{\widehat{p}_n} - H_{p_0}\|_\infty = O_p(1). \tag{2.9}$$

2.4. The support of the convex LSE

In the sequel, we will assume that p_0 has a finite support and we denote the support by $\{0, 1, \ldots, S\}$. Note that $S + 1$ is the last knot of p_0 in the sense that $\Delta p_0(S + 1) = p_0(S) > 0$ and $\Delta p_0(k) = 0$ for all integers $k > S + 1$. Under this assumption, it is natural to ask whether the support of $\widehat{p}_n$ is also finite. It turns out that the answer is affirmative as we now show in the following proposition.

Proposition 2.6. *If p_0 is supported on $\{0, \ldots, S\}$ with $S \in \mathbb{N} \setminus \{0\}$, then with probability one there exists n_0 such that for all $n \geq n_0$, the support of the LSE $\widehat{p}_n$ is either $\{0, \ldots, S\}$ or $\{0, \ldots, S+1\}$.*

3. Asymptotics of the convex LSE

In this section, we derive the weak limit of the LSE when the true distribution is supported on a finite set. The limit distribution is given in Section 3.1 in the most general setting where we do not make any additional assumption on the structure of the knots of the true p.m.f. p_0. It turns out that the limit distribution of the estimator involves all knots of p_0, see Theorem 3.2 below. This seems to contrast with the continuous case, where the limiting distribution of the LSE at a point depends only on the density (and its derivatives) of the observations at this point, so that the limit distribution is "localized". Moreover, the limit distribution of the MLE of a discrete log-concave p.m.f. given in [2] is localized in some sense. For these reasons, we provide in Section 3.3 below general characterizing conditions for such localizations to occur for the LSE of a discrete convex p.m.f. This comprehensive study of possible localization led us to find an error in the proof of Proposition 3 in [2], and to conclude that the limit distribution given in Theorem 5 of that paper is not correct.

3.1. The general setting

Assume that p_0 is supported on $\{0, \ldots, S\}$ with an unknown integer $S > 0$. From Proposition 2.6, it follows that with probability one, $\widehat{p}_n$ is supported on $\{0, \ldots, S+1\}$ provided that n is sufficiently large. Therefore, we consider the weak limit of $\widehat{p}_n$ on $\{0, \ldots, S+1\}$. To this end, first consider the weak limit of the empirical p.m.f. p_n. By standard results, the empirical process $\sqrt{n}(F_{p_n} - F_{p_0})$ weakly converges to $\mathbb{U}(F_{p_0})$ on $\{0, \ldots, S+1\}$, where $\mathbb{U}$ denotes a standard Brownian bridge from $(0, 0)$ to $(1, 0)$. With

$$\mathbb{W}(k) = \mathbb{U}\big(F_{p_0}(k)\big) - \mathbb{U}\big(F_{p_0}(k-1)\big) \tag{3.1}$$

for all integers $k = 0, \ldots, S+1$, we conclude that $\sqrt{n}(p_n - p_0)$ weakly converges to $\mathbb{W}$ on $\{0, \ldots, S+1\}$. Since $\widehat{p}_n$ is the minimizer of a criterion that involves p_n, its weak limit depends on $\mathbb{W}$. Theorem 3.2 below proves that the limiting distribution of the LSE is that of the minimizer (whose existence is proved in Theorem 3.1 below) of the criterion

$$\Phi(g) = \frac{1}{2}\sum_{k=0}^{S+1}\big(g(k) - \mathbb{W}(k)\big)^2 \tag{3.2}$$

over the set $\mathcal{C}(\mathcal{K})$ that we define now. Let $\mathcal{K}$ be the set of all interior knots of p_0, that is all knots of p_0 in $\{1, \ldots, S\}$. If p_0 does not have any interior knot (which means that p_0 is triangular p.m.f.), then $\mathcal{K} = \varnothing$. Associated with $\mathcal{K}$ is the following class of functions

$$\mathcal{C}(\mathcal{K}) = \big\{g = \big(g(0), \ldots, g(S+1)\big) \in \mathbb{R}^{S+2} \text{ such that } \Delta g(k) \geq 0 \text{ for all } k \in \{1, \ldots, S\} \setminus \mathcal{K}\big\}.$$

This means that $g \in \mathcal{C}(\mathcal{K})$ is convex between two successive knots of p_0. Define

$$\mathbb{H}(z) = \sum_{k=0}^{z-1}\mathbb{U}\big(F_{p_0}(k)\big) = \sum_{k=0}^{z-1}\sum_{j=0}^{k}\mathbb{W}(j) \tag{3.3}$$

for all $z \in \mathbb{N}$, with $\mathbb{H}(0) = 0$. Then, we have the following theorem.

Theorem 3.1. *The criterion* (3.2) *admits a unique minimizer* $\widehat{g}$ *over* $\mathcal{C}(\mathcal{K})$. *Furthermore, with probability one, an element* $\widehat{g} \in \mathcal{C}(\mathcal{K})$ *is the minimizer if and only if the process* $\widehat{\mathbb{H}}$ *defined on* $\{0, \ldots, S+2\}$ *by*

$$\widehat{\mathbb{H}}(x) = \sum_{k=0}^{x-1}\sum_{j=0}^{k}\widehat{g}(j) \tag{3.4}$$

with the convention that $\widehat{\mathbb{H}}(0) = 0$, *satisfies*

$$\widehat{\mathbb{H}}(x)\begin{cases} \geq \mathbb{H}(x), & \text{for all } x \in \{0, \ldots, S+2\}, \\ = \mathbb{H}(x), & \text{if } x \in \mathcal{K} \cup \{0, S+1, S+2\} \\ & \text{or } x \in \{1, \ldots, S\} \setminus \mathcal{K} \text{ satisfies } \Delta\widehat{g}(x) > 0. \end{cases} \tag{3.5}$$

In the above theorem, note that it is implicit that the minimizer $\widehat{g}$ is actually $\widehat{g}(\omega)$. We are now ready to establish the weak convergence of $\widehat{p}_n$. For $x \in \{0, \ldots, S+1\}$ define

$$\widehat{\mathbb{G}}(x) = \sum_{k=0}^{x} \widehat{g}(k) = \widehat{\mathbb{H}}(x+1) - \widehat{\mathbb{H}}(x). \tag{3.6}$$

Theorem 3.2. *If $\widehat{p}_n$ is the convex LSE of the true p.m.f. p_0 with support $\{0, \ldots, S\}$, then we have the joint weak convergence on $\{0, \ldots, S+1\}$ as $n \to \infty$:*

$$\begin{pmatrix} \sqrt{n}(H_{\widehat{p}_n} - H_{p_0}) \\ \sqrt{n}(F_{\widehat{p}_n} - F_{p_0}) \\ \sqrt{n}(\widehat{p}_n - p_0) \end{pmatrix} \Rightarrow \begin{pmatrix} \widehat{\mathbb{H}} \\ \widehat{\mathbb{G}} \\ \widehat{g} \end{pmatrix}.$$

3.2. Estimating the limiting distribution

The asymptotic distribution $\widehat{g}$ derived in Theorem 3.2 depends on the projection of a Gaussian vector, whose dispersion matrix depends on p_0, onto a set $\mathcal{C}(\mathcal{K})$ that depends on the interior knots of p_0. Since those knots are typically unknown, the asymptotic distribution cannot be directly used to build confidence intervals. Below, we fill the gap between the theoretical result of Theorem 3.2 and the practical problem of building such intervals. This is achieved via the construction of a random vector $\widehat{g}_n$ that weakly converges to $\widehat{g}$ without depending on some unknown parameter. The distribution of $\widehat{g}_n$ can easily be approximated via Monte-Carlo simulations, and can therefore be used to approximate the distribution of $\sqrt{n}(\widehat{p}_n - p_0)$.

To define $\widehat{g}_n$, let $S_n = \max\{X_1, \ldots, X_n\}$, and g_n be the centered Gaussian vector of dimension S_n+2 and whose dispersion matrix is given by Γ_n, the $(S_n+2) \times (S_n+2)$ matrix with component $(i+1, j+1)$ equal to $p_n(i)(1-p_n(i))$ for all $i = j$ and $-p_n(i)p_n(j)$ for all $i \neq j$, with $i, j = 0, \ldots, S_n+1$. Since g_n converges weakly to $\mathbb{W}$, by analogy to how $\widehat{g}$ was defined, it is natural to define $\widehat{g}_n$ as the minimizer of

$$\Phi_n(g) = \frac{1}{2} \sum_{k=0}^{S_n+1} \big(g(k) - g_n(k)\big)^2 \tag{3.7}$$

over a set $\mathcal{C}_n$ that approaches $\mathcal{C}(\mathcal{K})$ as $n \to \infty$. We feel it is useful to point out that defining $\mathcal{C}_n$ as the set of all functions $(g(0), \ldots, g(S_n+1)) \in \mathbb{R}^{S_n+2}$ such that $\Delta g(k) \geq 0$ for all $k \in \{1, \ldots, S_n\}$ with possible exception at the knots of $\widehat{p}_n$ does not work. Indeed, due to the fact that (with probability one, asymptotically) the set of all knots of $\widehat{p}_n$ contains the set of all knots of p_0 with typically a strict inclusion, the proposed set does not approaches $\mathcal{C}(\mathcal{K})$ as $n \to \infty$. An appropriate choice for $\mathcal{C}_n$ is given in the following theorem. In dealing with convergences, we view $X_1, \ldots, X_n$ as the first n terms of an infinite sequence $(X_i)_{i \in \mathbb{N}}$ of i.i.d. random variables.

Theorem 3.3. *Let $(v_n)_{n \in \mathbb{N}}$ be a sequence of positive numbers that satisfy*

$$\lim_{n \to \infty} v_n = 0 \quad \textit{and} \quad \lim_{n \to \infty} \sqrt{n} v_n = \infty. \tag{3.8}$$

Define g_n as above and let $\widehat{g}_n$ be the minimizer of (3.7) over the set $\mathcal{C}_n$ of all functions $g = (g(0), \ldots, g(S_n + 1)) \in \mathbb{R}^{S_n+2}$ such that $\Delta g(x) \geq 0$ for all $x \in \{1, \ldots, S_n\}$ with possible exceptions at the points x that satisfy $\Delta \widehat{p}_n(x) > v_n$. Then, conditionally on $(X_i)_{i\in\mathbb{N}}$, the random vector $\widehat{g}_n$ converges in distribution to $\widehat{g}$ in probability as $n \to \infty$.

Remark 3.4. The meaning of the above convergence should be clarified: $\widehat{g}_n$ is a random vector in $\mathbb{R}^{S_n+2}$ whereas $\widehat{g}$ is in $\mathbb{R}^{S+2}$. But since $P(S_n \leq S) = 1$, the convergence in distribution has a full meaning once $\widehat{g}_n$ is identified with the $(S + 2)$-dimensional vector $(\widehat{g}_n, 0, \ldots, 0)$, where the number of null entries is exactly $S - S_n$.

Remark 3.5. Note that in the inequality $\Delta \widehat{p}_n(x) > v_n$, $\widehat{p}_n$ can be replaced by p_n. The arguments used to show Theorem 3.3 will remain almost unchanged since we can use the central limit theorem. The same thing applies for the estimator Σ_n, which can be defined by taking any consistent estimator of p_0 for the purpose of approaching the Gaussian distribution of $\mathbb{W}$.

In practice, Theorem 3.3 can be used to simulate an approximation of the distribution of $\sqrt{n}(\widehat{p}_n - p_0)$ as follows: once the empirical estimator has been computed from the observations $X_1, \ldots, X_n$, compute Γ_n and simulate B (a large positive integer) independent copies $g_{n,1}, \ldots, g_{n,B}$ from a centered multivariate Gaussian distribution with dispersion matrix Γ_n. Then, for $j = 1, \ldots, B$, compute $\widehat{g}_{n,j}$ as the minimizer of the criterion (3.7) where g_n is replaced by $g_{n,j}$ over the set $\mathcal{C}_n$ as defined in Theorem 3.3. From Theorem 3.3, it follows that the empirical distribution of $\widehat{g}_{n,j}$, $j = 1, \ldots, B$ approaches the limit distribution $\widehat{g}$ of $\sqrt{n}(\widehat{p}_n - p_0)$.

As a consequence of Theorem 3.3, $\widehat{g}_n$ converges in distribution to $\widehat{g}$ unconditionally. This fact is less useful for practical applications but easier to illustrate with simulations as shown below in Section 4.

3.3. Localization

It follows from Theorem 3.2 above that the limiting distribution of $\widehat{p}_n$ at a fixed point involves all knots of p_0 in the general case. This seems to contrast with the continuous case where the limiting distribution of the LSE is localized in the sense that it depends only on the true density (and its derivatives) at the fixed point. Hence, a natural question is whether the convex LSE could be also localized in the discrete case.

To draw a correct comparison between what happens in the discrete and continuous cases, one has to go back to the working assumptions under which the limiting distribution has been derived in the latter case. In [13], it is assumed that the true convex density f defined on $[0, \infty)$ is twice continuously differentiable in a small neighborhood of a fixed point $x_0 > 0$ such that $f''(x_0) > 0$. In particular, the density is strictly convex at x_0. Under this assumption, the limiting distribution only depends on $f(x_0)$ and $f''(x_0)$. In Theorem 3.2, we do not consider any particular configuration for the knots of p_0. For the sake of comparison, and if we translate for the moment strict convexity at a point s in the support of p_0 as having s to be a triple knot, that is $s - 1$, s and $s + 1$ are successive knots of p_0, then it follows from Proposition 2.4 and Proposition 2.1 that with probability one there exists n_0 large enough such that for all $n \geq n_0$

$$H_{\widehat{p}_n}(s-1) = H_{p_n}(s-1), \qquad H_{\widehat{p}_n}(s) = H_{p_n}(s) \quad \text{and} \quad H_{\widehat{p}_n}(s+1) = H_{p_n}(s+1).$$

This implies that $\widehat{p}_n(s) = p_n(s)$ and the limiting distribution is simply that of the Gaussian random variable $\mathcal{N}(0, p_0(s)(1 - p_0(s)))$. In this case, the limit of the LSE at the point s is completely localized in the sense that it is not influenced by the remaining knots of p_0. The identity $\widehat{p}_n(s) = p_n(s)$ shows even the stronger fact that the localization is actually happening at the level of the estimator itself.

It is conceivable that other configurations lead to some form of localization of the weak limit. We provide below general characterizing conditions for such localizations to occur. In fact, the LSE and its weak limit get localized either to the left or right at any knot of p_0 that is either followed or preceded by another knot of p_0. In such cases, the limit of the LSE can be described only in terms of knots of p_0 that are either to the left or right of that knot; see comments after Theorems 3.6 and 3.7. In the sequel, we shall use the same notation as in Section 3.1.

First, we consider the question of localizing "to the left" of a knot. This means that given an interior knot s of p_0, we wonder whether the restriction to $\{0, \ldots, s\}$ of the limiting $\widehat{g}$ is distributed as the minimizer of the left-localized criterion

$$\Phi^{\leq s}(g) = \frac{1}{2} \sum_{k=0}^{s} \big(g(k) - \mathbb{W}(k)\big)^2$$

over the set

$$\mathcal{C}^{\leq s}(\mathcal{K}) = \big\{ g = \big(g(0), \ldots, g(s)\big) \in \mathbb{R}^{s+1} \text{ such that } \Delta g(k) \geq 0 \text{ for all } k \in \{1, \ldots, s-1\} \setminus \mathcal{K} \big\}.$$

The following theorem provides a necessary and sufficient condition for the answer to be positive. It also gives a necessary and sufficient condition for the restriction of $\sqrt{n}(\widehat{p}_n - p_0)$ to $(0, \ldots, s)$ to converge to the left-localized minimizer.

Theorem 3.6. *Assume that the support of p_0 is finite. Then for an arbitrary interior knot s of p_0, there exists a unique minimizer of $\Phi^{\leq s}$ over $\mathcal{C}^{\leq s}(\mathcal{K})$. The minimizer is equal to $(\widehat{g}(0), \ldots, \widehat{g}(s))$ if, and only if,*

$$\widehat{\mathbb{G}}(s) = \mathbb{U}\big(F_{p_0}(s)\big). \tag{3.9}$$

Moreover, $\sqrt{n}(\widehat{p}_n(0) - p_0(0), \ldots, \widehat{p}_n(s) - p_0(s))$ converges in distribution to the minimizer of $\Phi^{\leq s}$ over $\mathcal{C}^{\leq s}(\mathcal{K})$ if, and only if,

$$F_{\widehat{p}_n}(s) = F_{p_n}(s) + o_p\big(n^{-1/2}\big). \tag{3.10}$$

Consider the case where s is a double knot, in the sense that both s and $s+1$ are knots of p_0. It follows from Proposition 2.4 together with the characterization in Proposition 2.1, that with probability one, both $H_{\widehat{p}_n}(s) = H_{p_n}(s)$ and $H_{\widehat{p}_n}(s+1) = H_{p_n}(s+1)$ hold true for sufficiently large n. Therefore, $F_{\widehat{p}_n}(s) = F_{p_n}(s)$ with probability one for sufficiently large n, so that (3.10) holds and the limiting distribution is left-localized.

Now, we consider the question of localizing "to the right" of a knot. This means that given an interior knot s of p_0, we wonder whether the restriction to $\{s, \ldots, S+1\}$ of the limiting $\widehat{g}$ is

distributed as the minimizer of the right-localized criterion

$$\Phi^{\geq s}(g) = \frac{1}{2}\sum_{k=s}^{S+1}\big(g(k) - \mathbb{W}(k)\big)^2$$

over the set $\mathcal{C}^{\geq s}(\mathcal{K})$ of all $g = (g(s), \ldots, g(S+1)) \in \mathbb{R}^{S-s}$ such that $\Delta g(k) \geq 0$ for all $k \in \{s+1, \ldots, S\} \setminus \mathcal{K}$. A necessary and sufficient condition for the answer to be positive, is given below.

Theorem 3.7. *Assume that the support of p_0 is finite. Then for an arbitrary interior knot s of p_0, there exists a unique minimizer of $\Phi^{\geq s}$ over $\mathcal{C}^{\geq s}(\mathcal{K})$. The minimizer is equal to $(\widehat{g}(s), \ldots, \widehat{g}(S+1))$ if, and only if,*

$$\widehat{\mathbb{G}}(s-1) = \mathbb{U}\big(F_{p_0}(s-1)\big). \tag{3.11}$$

Moreover, $\sqrt{n}(\widehat{p}_n(s) - p_0(s), \ldots, \widehat{p}_n(S+1) - p_0(S+1))$ converges in distribution to the minimizer of $\Phi^{\geq s}$ over $\mathcal{C}^{\geq s}(\mathcal{K})$ if, and only if,

$$F_{\widehat{p}_n}(s-1) = F_{p_n}(s-1) + o_p\big(n^{-1/2}\big). \tag{3.12}$$

Similar as above, (3.12) holds in the specific case where both $s-1$ and s are knots of p_0.

4. Numerical aspects

4.1. A Dykstra algorithm for computing the asymptotic distribution

In monotone or concave/convex regression, active set methods are often proposed to compute the estimators; see for example [15] or [14] for more recent work. One may also refer to [18] for a more general method of non parametric regression under a shape constraint.

Here, we describe a simple algorithm that enables us to simulate a sample of any size from the asymptotic distribution $\widehat{g}$ of $\sqrt{n}(\widehat{p}_n - p_0)$ in the case where p_0 has a finite unknown support $\{0, \ldots, S\}$, see Theorem 3.2. The algorithm also enables us to simulate a sample of any size from the conditional distribution of $\widehat{g}_n$ given in Theorem 3.3, see Remark 4.1 below. Let $s_1 < \cdots < s_m$ be the interior knots of p_0, and put $s_0 = 0$ and $s_{m+1} = S+1$. Then, define

$$\mathcal{C}_j = \big\{g = \big(g(0), \ldots, g(S+1)\big) \in \mathbb{R}^{S+2} \text{ such that } \Delta g(k) \geq 0 \text{ for all } k \in \{s_j, \ldots, s_{j+1}\}\big\}$$

for $j = 0, \ldots, m$. By definition (see Theorem 3.1), $\widehat{g}$ is the unique minimizer of the criterion Φ in (3.2) over $\mathcal{C}(\mathcal{K})$, which means that $\widehat{g}$ can be viewed as the projection of $\mathbb{W}$ onto $\bigcap_{j=0}^{m}\mathcal{C}_j$. Since the $\mathcal{C}_j$'s are closed convex cones, the solution can be found using the algorithm of [11] which proceeds by performing cyclic projections onto the convex cones $\mathcal{C}_0, \ldots, \mathcal{C}_m$. Although details of the algorithm are given in [11], we describe here how these projections are performed.

Set $g^{(0)} \equiv (\mathbb{W}(0), \mathbb{W}(1), \ldots, \mathbb{W}(S+1))$ and $u_j^{(0)} = 0$ for $j = 0, \ldots, m$. Then, for $i \geq 1$, iterate the three following steps.

(1) Compute $g_j^{(i)}$, the projection of $g_{j-1}^{(i)} - u_j^{(i-1)}$ onto $\mathcal{C}_j$ for $j = 0, \ldots, m$.
(2) Set $u_j^{(i)} \equiv g_j^{(i)} - (g_{j-1}^{(i)} - u_j^{(i-1)})$.
(3) Set $i = i + 1$ and go to (1).

Granted that we know how to obtain the convex projections $g_j^{(n)}$, convergence of the above algorithm is a consequence of Theorem 3.1 of [11]. The projection onto each cone $\mathcal{C}_j$, $j = 0, \ldots, m$ can be efficiently computed using the R function `conreg` available in the R package *COBS*; see [19] for more details.

Now, for any fixed integer $N \geq 1$, a sample of size N from the same distribution as $(\widehat{g}(0), \ldots, \widehat{g}(S+1))$ can be done as follows: we generate a centered Gaussian vector $(W_0, \ldots, W_S)$ whose dispersion matrix is given by Γ_0, the $(S+1) \times (S+1)$ matrix with component $(i+1, j+1)$ equal to $p_0(i)(1 - p_0(i))$ for all $i = j$ and $-p_0(i)p_0(j)$ for all $i \neq j$, with $i, j = 0, \ldots, S$. This can be done using the R function `rmvnorm` available from the *mvtnorm* package. In the second step, we compute the piecewise convex projection of $(W_0, \ldots, W_S, 0)$ as described above, and the two steps are then repeated N times.

Remark 4.1. Let (v_n) be a sequence satisfying (3.8). Once $X_1, \ldots, X_n$ have been observed, a sample from the conditional distribution of $\widehat{g}_n$ given in Theorem 3.3 can be simulated by using the above algorithm with $s_1, \ldots, s_m$ replaced by the points x that satisfy $\Delta \widehat{p}_n(x) > v_n$, $s_{m+1} = S_n + 1$ where we recall that $S_n = \max\{X_1, \ldots, X_n\}$, and $g^{(0)}$ replaced by a centered Gaussian vector in $\mathbb{R}^{S_n+2}$ with dispersion matrix Γ_n. Alternatively, according to Remark 3.5, the points $s_1, \ldots, s_m$ might be replaced by the points x that satisfy $\Delta p_n(x) > v_n$.

4.2. How well the true knots are captured

Recall that Proposition 2.4 implies that with probability one, having enough large sample sizes ensures that a knot of the true p.m.f. p_0 is also a knot of the LSE $\widehat{p}_n$. However, the proposition does not indicate how large n should be. To gain some insight into the relationship between the size of the sample at hand and whether the knots of the estimator include all true knots, we have carried out a simulation study with samples of size $n \in \{50, 200, 800, 3200, 12\,800, 51\,200\}$. Given a simulated sample of size n from a distribution p_0, the convex LSE $\widehat{p}_n$ was computed using the algorithm described in [8].

To define the convex p.m.f.'s under which the samples were generated, we use the fact that a p.m.f. p_0 is convex if and only if p_0 admits the mixture representation

$$p_0 = \sum_{j \geq 1} \pi_j T_j, \tag{4.1}$$

where T_j is the triangular distribution defined by (2.7), $0 \leq \pi_j \leq 1$ and $\sum_{j \geq 1} \pi_j = 1$, see Theorem 7 in [8]. The representation is unique and the mixing weights are given by

$$\pi_j = \frac{j(j+1)}{2} \Delta p(j)$$

Table 1. Mixing weights π_j for the convex p.m.f.'s $p_1, p_2, p_3, p_4, p_5, p_6$

p.m.f.	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}
p_1	0	0	0	2/3	0	0	0	0	0	0	1/3
p_2	1/3	0	0	0	0	1/2	0	0	0	0	1/6
p_3	0	1/6	0	0	1/6	0	0	0	1/2	0	1/6
p_4	0	0	0	1/6	0	1/6	0	1/12	0	1/2	1/12
p_5	0	0	1/6	1/12	1/4	0	1/12	0	1/6	1/6	1/6
p_6	0	1/12	1/6	1/12	1/12	1/12	1/12	1/12	1/12	1/6	1/12

for $j \geq 1$. Note that in particular, j is a knot of p_0 if and only if $\pi_j > 0$, and if the support of p_0 takes the form $\{0, \ldots, S\}$, then $\pi_{S+1} > 0$ whereas $\pi_j = 0$ for all $j > S + 1$.

In our simulations, the samples were generated from four convex p.m.f.'s that are all supported on $\{0, \ldots, 10\}$. We give in Table 1 the values of the mixing weights $\pi_j, 1 \leq j \leq 11$ for various p.m.f.'s p_0 that are denoted by p_1, p_2, p_3 and p_4.

Figure 1 shows the process $\sqrt{n}(H_{\widehat{p}_n} - H_{p_n})$ together with the knots of the LSE $\widehat{p}_n$ and the true knots, for a sample of size $n \in \{50, 200\}$ generated from p_3. In these examples, it can be seen that in accordance with Proposition 2.1, $H_{\widehat{p}_n}(z) \geq H_{p_n}(z)$ with an equality at all knots z of $\widehat{p}_n$. However, the sample sizes are not large enough to ensure that the knots of $\widehat{p}_n$ include all knots of the true p.m.f. p_3. Neither they are large enough to ensure that the support of $\widehat{p}_n$ is included in $\{0, \ldots, S + 1\}$ where $S = 10$ denotes the greatest point in the support of the true p.m.f.; see Proposition 2.6. Figure 2 is similar to Figure 1 but now, the samples are generated from the triangular p.m.f. T_{11}, which means that the mixing probabilities are $\pi_{11} = 1$ and $\pi_j = 0$

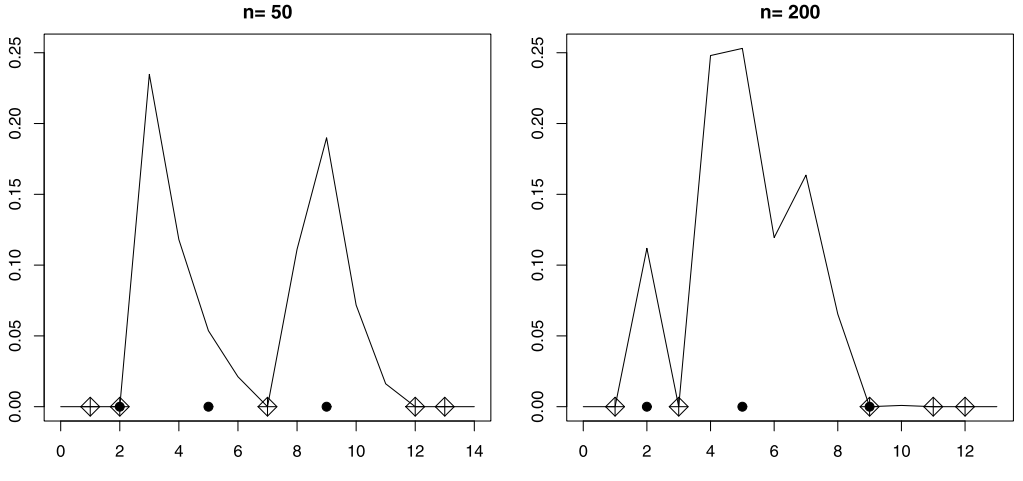

Figure 1. The figures show the process $\sqrt{n}(H_{\widehat{p}_n} - H_{p_n})$ for a random sample of size n as shown. The samples were generated from p_3 as in Table 1. The diamond symbols depict the knots of the LSE $\widehat{p}_n$ computed based on the samples, whereas the bullets show the locations of the true knots.

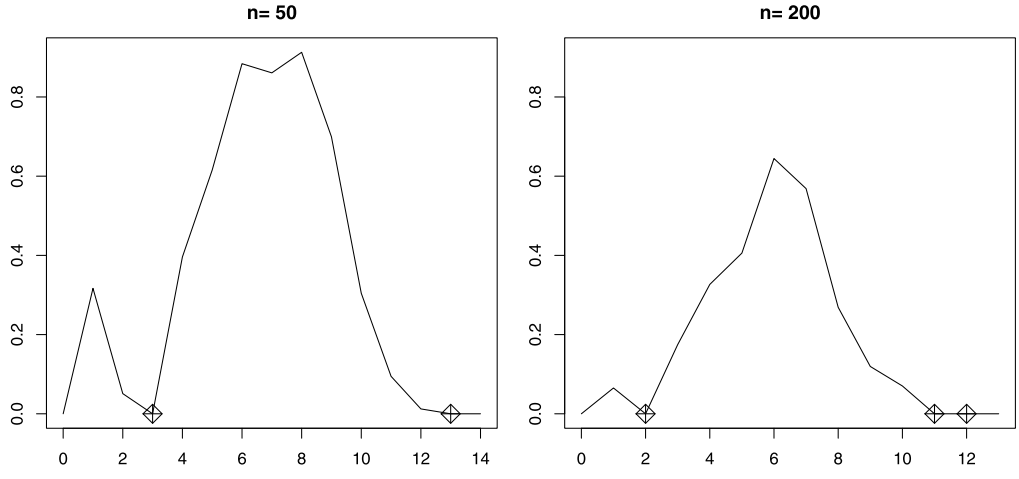

Figure 2. The figures show the process $\sqrt{n}(H_{\widehat{p}_n} - H_{p_n})$ for a random sample of size n as shown. The samples were generated from a triangular p.m.f. supported on $\{0, 1, \ldots, 10\}$. The diamond symbols depict the knots of the LSE $\widehat{p}_n$ computed based on the samples.

for all $j \neq 11$. Again, we observe that $H_{\widehat{p}_n}(z) \geq H_{p_n}(z)$ with an equality at all knots z of $\widehat{p}_n$. In the case of a sample size $n = 50$, the knots of $\widehat{p}_n$ do not include the only true knot 11 and the support of $\widehat{p}_n$ is not included in $\{0, \ldots, 11\}$. On the other hand, in the case of a larger sample size $n = 200$, the knots of $\widehat{p}_n$ include the only true knot 11 and $\widehat{p}_n$ is supported on $\{0, \ldots, 11\}$.

Figures 1 and 2 are not sufficient to gain full insight into the connection between the knots of $\widehat{p}_n$ and the true knots since only one sample is considered in each situation. Thus, for each considered sample size and distribution, we simulated independently 1000 samples to evaluate the probability that the knots of $\widehat{p}_n$ include all true knots. The probabilities are estimated by empirical frequencies. Results are reported in Table 2. As expected, the empirical frequency increases as n increases. It is typically larger in cases of true distributions with only few knots than in cases of true distributions with many knots.

4.3. Assessing the convergence of the convex LSE to the weak limit

To assess convergence of the estimation error to the right weak limit, consider $\widehat{\mathbb{F}}^{(j)}_{n,M}$ and $\mathbb{F}^{(j)}_{M'}$ to be respectively the empirical distributions of $\sqrt{n}(\widehat{p}_n(j) - p_0(j))$ and $\widehat{g}(j)$ for $j \in \{0, \ldots, S+1\}$ based on M and M' independent replications. Here, M and M' will be chosen to be large. More explicitly, a sample of n independent random variables $X^{(i)}_1, \ldots, X^{(i)}_n$ is drawn from p_0 for each $i = 0, \ldots, M$ to form a sample of size M from the distribution of the estimation error. Note that this sample is multidimensional of dimension $S + 1$ hence our need to consider the marginal components of its distribution. Similarly, we draw a sample of size M' from the distribution of

Table 2. Empirical frequencies in % of having all knots of the true convex p.m.f. among those of the estimator $\widehat{p}_n$ for $n \in \{50, 200, 800, 3200, 12\,800, 51\,200\}$. The empirical frequencies are based on 1000 replications for each sample size and distribution. The true convex p.m.f.'s, p_1, p_2, p_3, p_4, p_5 and p_6 have 1 and 2, 3, 4, 6 and 9 interior knots respectively. See text for the exact expressions of those p.m.f.'s

n	p_1	p_2	p_3	p_4	p_5	p_6
50	73.9	44.8	4.2	0.0	0.0	0.0
200	96.8	82.8	14.5	2.1	0.3	0.0
800	100	99.5	47.3	9.0	3.7	0.0
3200	100	100	84.1	31.4	32.4	4.3
12 800	100	100	99.1	66.3	71.2	33.1
51 200	100	100	100	92.8	95.9	88.5

weak limit using the algorithm described in Section 4.1. Define now

$$D_{n,M,M'} = \sup_{0 \le j \le S+1} \left\| \widehat{\mathbb{F}}^{(j)}_{n,M} - \mathbb{F}^{(j)}_{M'} \right\|_\infty.$$

We will use this random variable to assess the established convergence. This is based on the fact that it is expected to become small for large n. Since the "target" distributions are $\mathbb{F}^{(j)}_{M'}$, $j = 0, \ldots, S+1$, we choose $M' > M$. Also, we store those obtained empirical distributions and reuse them while sampling many times from the estimation error. This enables us to obtain independent realizations from $\widehat{\mathbb{F}}^{(j)}_{n,M}$ while $\mathbb{F}^{(j)}_{M'}$, $j = 0, \ldots, S+1$ are fixed. To visualize the statistical summary of $D_{n,M,M'}$, the obtained outcomes are represented in the form of boxplots. Those were based on 100 replications of $\widehat{\mathbb{F}}^{(j)}_{n,M}$, $j = 0, \ldots, S+1$. Here, $M = 1000$, $M' = 5000$ and $n \in \{50, 100, 500, 1000, 5000, 10\,000, 50\,000\}$. In the simulations, we have taken the following true convex p.m.f.'s which are all supported on $\{0, \ldots, 10\}$:

- The triangular p.m.f. p_0 given by $p_0(i) = (11 - i)_+/66$ for $i \in \mathbb{N}$.
- The convex p.m.f.'s p_1, p_2, p_3, p_4, p_5, p_6 considered above in Section 4.2.
- The p.m.f., p_7, of a truncated Geometric p.m.f. with success probability equal to $1/2$, given by $p_7(i) = (1 - 2^{-11})^{-1} 2^{-(i+1)}$ for $i \in \{0, \ldots, 10\}$ and $p_7(i) = 0$ otherwise.

Note that if a geometric p.m.f. is always convex on $\mathbb{N}$, this is not the case anymore after truncation. Indeed, convexity of the latter version holds true if and only if the waiting probability is $\le 1/2$. A simple proof of this fact can be found in Section 5.

Lemma 4.2. *Let S be a positive integer and $q \in (0, 1)$. The truncated Geometric distribution, defined by*

$$p(i) = \frac{q^i (1 - q)}{1 - q^{S+1}}, \qquad i \in \{0, \ldots, S\}$$

and $p(i) = 0$ for all integers $i \ge S + 1$, is convex if and only if $q \le 1/2$.

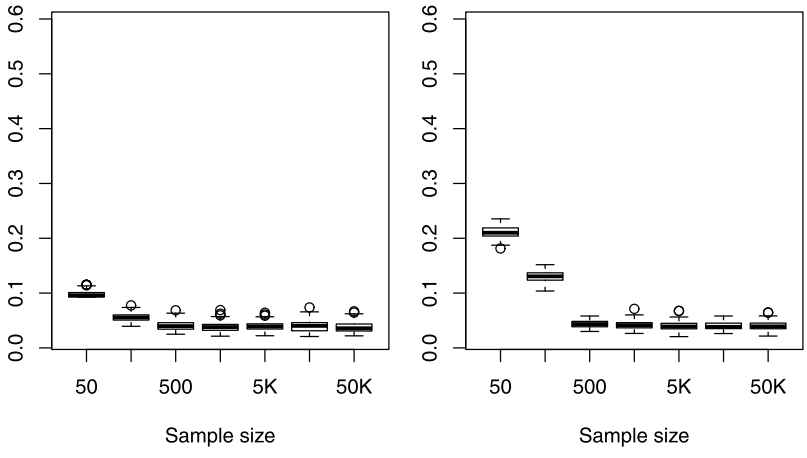

Figure 3. Boxplots of $D_{n,M,M'}$ with $M = 1000$ and $M' = 5000$. The sample size n is as indicated and the true convex p.m.f. is p_0 on the left figure, and p_1 on the right figure. See text for details.

To compute the uniform distance between $\widehat{\mathbb{F}}^{(j)}_{n,M}$ and $\mathbb{F}^{(j)}_{M'}$, we computed the maximal value of the absolute difference on a discretized grid $\{-4, -3.99, \ldots, 3.99, 4\}$ with a regular step equal to 0.01. The boxplots shown in Figures 3–6 give support to the asymptotic theory of the estimation error of the LSE in case the true p.m.f. is one of the selected convex p.m.f.'s $p_0, p_1, \ldots, p_7$. Interestingly, weak convergence seems not to happen at the same speed. For the triangular p.m.f. p_0, the boxplots appear to stabilize for $n \geq 1000$ whereas the obtained boxplots for the other distributions seem to indicate that convergence has not been yet attained. According to our nu-

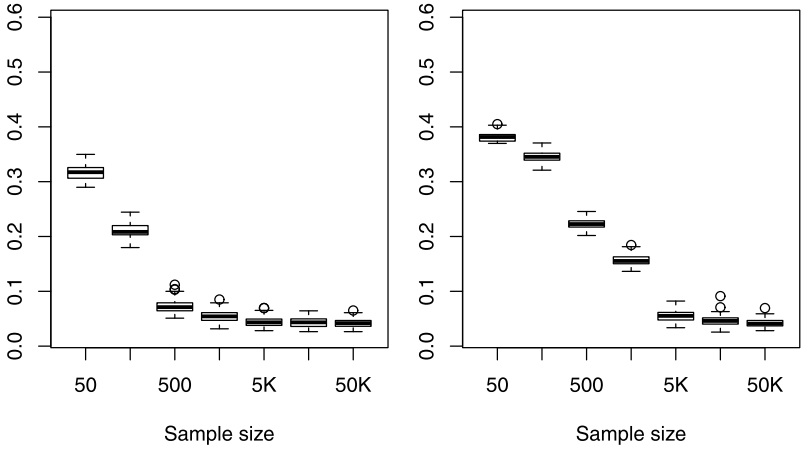

Figure 4. Boxplots of $D_{n,M,M'}$ with $M = 1000$ and $M' = 5000$. The sample size n is as indicated and the true convex p.m.f. is p_2 on the left figure, and p_3 on the right figure. See text for details.

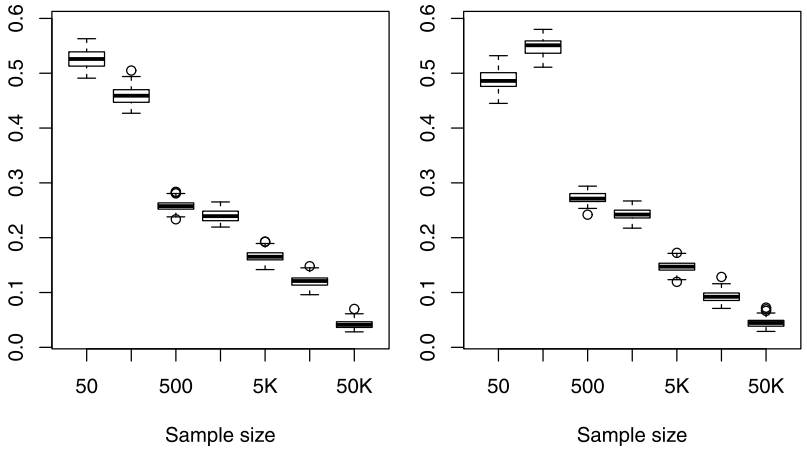

Figure 5. Boxplots of $D_{n,M,M'}$ with $M = 1000$ and $M' = 5000$. The sample size n is as indicated and the true convex p.m.f. is p_4 on the left figure, and p_5 on the right figure. See text for details.

merical findings in Section 4.2, large sample sizes could be required for the estimator to be able to capture these knots. Thus, the slow convergence to the true limit for $p_i, 1 \leq i \leq 7$ could be partially explained by the fact that those p.m.f.'s have all interior knots, as opposed the relatively fast convergence in the case of the triangular p.m.f. p_0 which has none. We would like to note that all points in $\{1, \ldots, 10\}$ are interior knots of the truncated geometric p.m.f. p_7 since it is strictly convex of its support. Hence, it is the p.m.f. with the largest number of interior knots

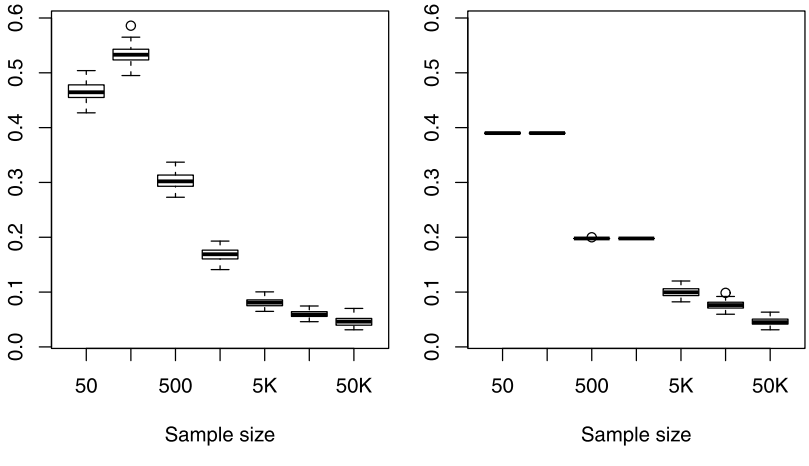

Figure 6. Boxplots of $D_{n,M,M'}$ with $M = 1000$ and $M' = 5000$. The sample size n is as indicated and the true convex p.m.f. is p_6 on the left figure, and the truncated geometric, p_7, with success probability equal to $1/2$. See text for details.

and also the one for which the convergence seems to be the slowest. Also, the limit $\widehat{g}$ reduces in this particular case to $\mathbb{W}$. Indeed, the empirical p.m.f. p_n is convex for large sample sizes (see Proposition 2.4) so that $\widehat{p}_n = p_n$, and hence $\sqrt{n}(\widehat{p}_n - p_7)$ has the same limit distribution $\mathbb{W}$ as $\sqrt{n}(p_n - p_7)$. Another way of viewing this is to note that the required convexity of the minimizer $\widehat{g}$ between the knots becomes a superfluous constraint since it is always satisfied by the straight line connecting $\mathbb{W}$ at two given knots. This is true only in this case because all knots are consecutive.

We would like to finish this section by adding that it is of course impossible to have a precise statement about the speed of convergence in case the true convex p.m.f. is known to be finitely supported. However, our numerical findings indicate that in applications such as construction of asymptotic confidence bands, one has to keep in mind that moderate sample sizes may not be enough to obtain good coverage. Finally, note that in our assessment we have assumed that the distribution of $\widehat{g}$ is continuous. We believe this is true but we do not intend to prove it here as it is beyond the scope of this work.

4.4. Assessing how well the weak limit is estimated

We finish this numerical section by a small simulation study which aims at assessing how well the weak limit is estimated using out Theorem 3.3 above. Choosing the appropriate sequence $(v_n)_n$ among the infinitely many possibilities is not an easy task. One can easily see that if $\sqrt{n}v_n$ converges too quickly or too slowly to ∞, then one may need very big sample sizes to be able to finally observe knots at all or to extract the true knots from the total set of knots of the estimator (we know that this set will include random ones). To see how the theory works in practice, we have carried out a simulation study with the four different convex p.m.f.'s, p_0, p_1, p_2 and p_3, which were already defined above. We recall that these p.m.f.'s are all supported on $\{0, 1, \ldots, 10\}$ and have 0, 1, 2 and 3 interior knots respectively. To assess how well the (true) weak limit is estimated, we have computed the maximal value of the uniform distance between the marginal empirical distribution functions of the estimated weak limit as defined in Theorem 3.3 and that of the true weak limit based on $M = 1000$ and $M' = 5000$ replications, respectively. As in the section above, the uniform distance was approximated by computing the maximal value of the absolute difference on a discretized grid $\{-4, -3.99, \ldots, 3.99, 4\}$ with a regular step equal to 0.01. The sequence v_n was chosen to be equal to $v_n = \sqrt{\log(\log(n))/n}$. The other choices we have tried include $\log(n)/\sqrt{n}$, $\sqrt{\log(n)/n}$, $\log(\log(n))/\sqrt{n}$ and all of them performed worse that the selected v_n in the sense that larger samples sizes were needed before good convergence to the true weak limit was observed. Based on 100 replications and for each sample size $n \in \{500, 1000, 5000, 10\,000, 50\,000\}$, we have computed the boxplots of the approximated uniform distance between the empirical distributions. The results are shown in Figure 7 for the considered p.m.f.'s $p_i, i = 0, 1, 2, 3$. As expected, bigger sample sizes are needed to observe convergence to the true weak limit for p.m.f.'s with larger number of knots. It remains to know how the sequence $(v_n)_n$, which somehow plays a role similar to bandwidth in the context of kernel estimation, should be chosen. Ideally, such a choice should be data-based, but we do not tackle this question in this present work and leave it to a future investigation.

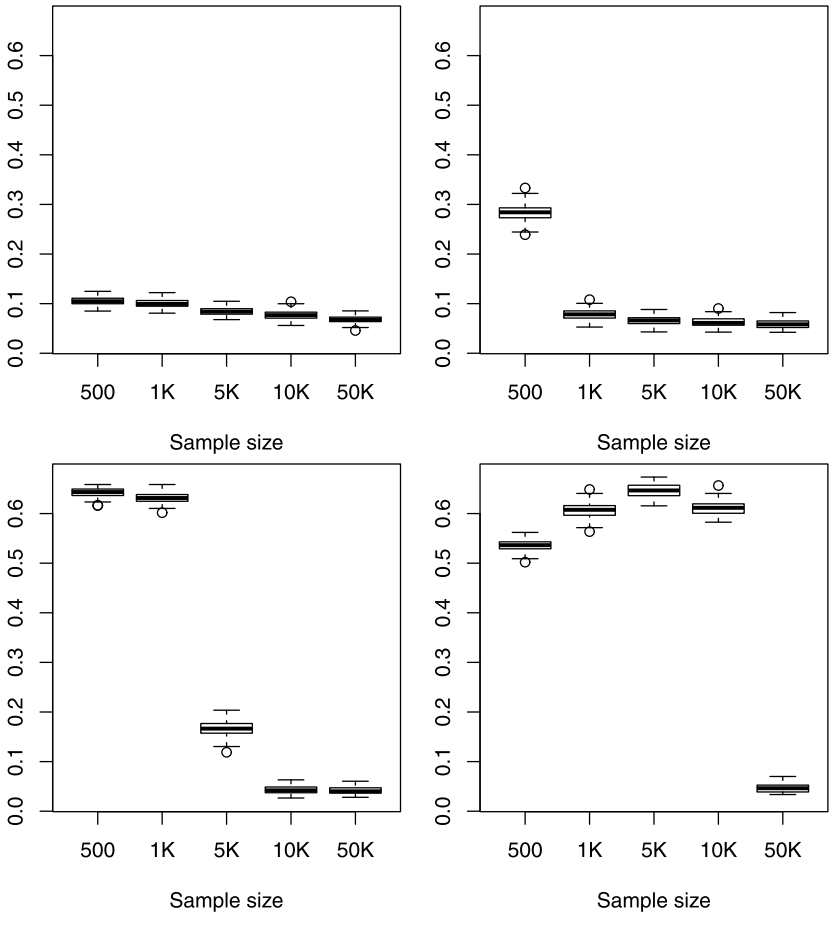

Figure 7. Boxplots of the maximal value between the marginal empirical distribution functions of the true and estimated weak limit with $v_n = \sqrt{\log(\log(n))/n}$, see text for further details. The sample size n is as indicated. The boxplots are based on 100 replications. The top ones correspond to the true convex p.m.f.'s p_0 (left) and p_1 (right) and the bottom ones to p_2 (left) and p_3 (right).

5. Proofs

Proof of Proposition 2.2. Assume we observe only $X_1 > 0$. We will show that the likelihood is maximized at a triangular p.m.f. To see this, define the function $f(x) = 2(x - X_1)/(x(x+1))$ for $x \in (X_1, \infty)$. Then, the first derivative of f is given by $f'(x) = 2g(x)/(x^2(x+1)^2)$ where $g(x) = -x^2 + 2xX_1 + X_1$. The function g is strictly concave on $\mathbb{R}$ with $g(X_1 + 1) > 0$ and $g(x) \to -\infty$ as $x \to \infty$. This means that there exists a unique $x_0 \in (X_1 + 1, \infty)$ such that $g(x_0) = 0$, $g(x) > 0$ for all $x \in (X_1 + 1, x_0)$ and $g(x) < 0$ for all $x \in (x_0, \infty)$. Hence, f achieves

its maximum over (X_1+1,∞) at the unique point x_0. This in turn implies that there exists an integer $j_0 \geq X_1+1$ such that $f(j_0) \geq f(j)$ for all integers $j \geq X_1+1$. Now, consider an arbitrary convex p.m.f. p on $\mathbb{N}$ and recall that p admits the mixture representation (4.1) where T_j is the triangular distribution defined by (2.7), $0 \leq \pi_j \leq 1$ and $\sum_{j\geq 1}\pi_j = 1$. Since

$$T_j(X_1) = \begin{cases} 0, & \text{for } j \leq X_1, \\ f(j), & \text{for } j \geq X_1+1, \end{cases}$$

where f was defined above, we conclude that for all convex p.m.f. p on $\mathbb{N}$, we have

$$T_{j_0}(X_1) \geq \sum_{j\geq 1}\pi_j T_j(X_1) = p(X_1). \tag{5.1}$$

Note now that in the case of a single observation X_1, the log-likelihood criterion reduces to

$$\ell_1(p) = \log\big(p(X_1)\big). \tag{5.2}$$

Combining the identities in (5.1) and (5.2) ensures existence of the MLE. Furthermore, the calculations above imply that there exists some integer $j_0 \geq X_1+1$ such that T_{j_0} is a solution. However, we shall see that it is not the only one when $X_1 > 1$.

To be able to exhibit an explicit value of j_0, we seek integers $j \geq X_1+1$ such that

$$\ell_1(T_{j+1}) - \ell_1(T_j) \leq 0 \quad \text{and} \quad \ell_1(T_j) - \ell_1(T_{j-1}) \geq 0$$

or equivalently (using again (5.2) and monotonicity of the logarithm), such that

$$T_{j+1}(X_1) - T_j(X_1) \leq 0 \quad \text{and} \quad T_j(X_1) - T_{j-1}(X_1) \geq 0.$$

This is in turn equivalent to

$$j(j+1-X_1) \leq (j+2)(j-X_1) \quad \text{and} \quad (j-1)(j-X_1) \geq (j+1)(j-1-X_1)$$

for $j \geq X_1+1$. These conditions are fulfilled if and only if $2X_1 \leq j \leq 2X_1+1$ with $j \geq X_1+1$. This implies that $j_0 \in \{2X_1, 2X_1+1\} \cap [X_1+1,\infty)$ where we recall that j_0 is such that $\ell_1(p)$ achieves its maximum over all convex p.m.f.'s p on $\mathbb{N}$ at T_{j_0}. Note that if $j_0 = X_1+1$, then we necessarily have $X_1 = 1$ (since $X_1 = 0$ was excluded). In this case, $j_0 = 2X_1 = 2$. If $X_1 > 1$, then $j_0 > X_1+1$ and the above calculations imply that $j_0 \in \{2X_1, 2X_1+1\}$. In this case, one can easily verify that

$$\ell_1(T_{2X_1}) = \ell_1(T_{2X_1+1}) = -\log(2X_1+1),$$

which is the maximal value of the log-likelihood. This also implies that the same maximal value is attained at any mixture $\pi T_{2X_1} + (1-\pi)T_{2X_1+1}$ with $\pi \in (0,1)$, giving rise to an infinite numbers of solutions.

We now prove that the LSE (which uniquely exists by strict convexity of the ℓ_2 norm) takes the form T_j for some j. The least squares criterion defined for one observation X_1 is

$$\begin{aligned}\Phi_1(T_j) &= \frac{1}{2}\sum_{i=0}^{j-1} T_j(i)^2 - T_j(X_1) + \frac{1}{2} \\ &= \frac{2j+1}{3j(j+1)} - \frac{2(j-X_1)_+}{j(j+1)} + \frac{1}{2}.\end{aligned}$$

It follows from Theorem 1 in [8] that the greatest support point of the LSE is greater than or equal X_1. Since the greatest support point of T_j is $j-1$, we restrict attention to integers $j \geq X_1 + 1$ such that the following conditions hold

$$\Phi_1(T_{j+1}) - \Phi_1(T_j) \geq 0 \quad \text{and} \quad \Phi_1(T_j) - \Phi_1(T_{j-1}) \leq 0. \tag{5.3}$$

A straightforward calculation shows that

$$\Phi_1(T_{j+1}) - \Phi_1(T_j) = \frac{4j-2-12X_1}{3j(j+1)(j+2)} \quad \text{and} \quad \Phi_1(T_j) - \Phi_1(T_{j-1}) = \frac{4j-6-12X_1}{3j(j-1)(j+1)}.$$

Then, conditions in (5.3) hold if and only if

$$4j - 2 - 12X_1 \geq 0 \quad \text{and} \quad 4j - 6 - 12X_1 \leq 0,$$

which is equivalent to $\frac{1}{2} + 3X_1 \leq j \leq \frac{3}{2} + 3X_1$ with $j \in \mathbb{N}$. Then the unique integer $j \geq X_1 + 1$ satisfying (5.3) is $3X_1 + 1$. Now, using the notation in (2.4) and (2.5), we have for all $i \geq 0$ and $j \geq 1$

$$F_{p_n}(i) = \begin{cases} 0, & \text{if } i < X_1, \\ 1, & \text{if } i \geq X_1; \end{cases}$$

$$H_{p_n}(i) = \begin{cases} 0, & \text{if } i < X_1 + 1, \\ i - X_1, & \text{if } i \geq X_1 + 1; \end{cases}$$

and

$$F_{T_j}(i) = \begin{cases} \dfrac{(2j-1)i - i^2 + 2j}{j(j+1)}, & \text{if } i \leq j-1, \\ 1, & \text{if } i \geq j; \end{cases}$$

$$H_{T_j}(i) = \begin{cases} \dfrac{i(i-1)}{6j(j+1)}\left(6j - 2i - 2 + \dfrac{12j}{i-1}\right), & \text{if } i \leq j, \\ \dfrac{1}{3}(2j+1) + (i-j), & \text{if } i \geq j+1. \end{cases}$$

It is not difficult to check that $H_{T_j}(j) = H_{p_n}(j)$ if and only if $j = 3X_1 + 1$. Moreover, we have $H_{T_{3X_1+1}}(i) \geq H_{p_n}(i)$ for all $i \geq 0$. The later holds because $H_{p_n}(i) = 0$ for $i < X_1 + 1$,

$H_{T_{3X_1+1}}(i) = H_{p_n}(i)$ for all integers i such that $i > 3X_1 + 1$ and the inequality also holds for $X_1 + 1 \le i \le 3X_1 + 1$ since the sequence $h(i) = H_{T_{3X_1+1}}(i) - H_{p_n}(i)$ decreases on set $\{X_1 + 1, \ldots, 3X_1 + 1\}$ with $h(3X_1 + 1) = 0$. Since j is the only knot of T_j, it then follows from the characterization of the LSE proved in Proposition 2.1 that the LSE is equal to T_{3X_1+1}. □

Proof of Proposition 2.3. For all sequences q, we have that $\|q\|_2 \le \|q\|_\infty^{1/2}\|q\|_1^{1/2}$ whence

$$\begin{aligned}\|p_n - p_0\|_2 &\le \|p_n - p_0\|_\infty^{1/2}\|p_n - p_0\|_1^{1/2}\\ &\le \|p_n - p_0\|_\infty^{1/2} \to 0\end{aligned}$$

almost surely by the Glivenko–Cantelli theorem. Now, for all sequences q and $r \in [2, \infty]$, one has $\|q\|_r \le \|q\|_2$, so it follows from Theorem 4 of [8] that

$$\begin{aligned}\|\widehat{p}_n - p_0\|_r &\le \|\widehat{p}_n - p_0\|_2\\ &\le \|p_n - p_0\|_2 \to 0, \qquad \text{a.s.}\end{aligned}$$

which completes the proof. □

Proof of Theorem 2.5. Theorem 6 of [8] implies (2.8). Let $\{0, \ldots, S\}$ denote again the support of p_0. Since $\widehat{p}_n$ is a proper p.m.f., it follows from Proposition 2.6 that $F_{\widehat{p}_n}(z) - F_{p_0}(z) = 0$ for $z \ge S + 1$. Hence, for all $z \in \mathbb{N}$ we have that

$$\begin{aligned}\sqrt{n}\big|F_{\widehat{p}_n}(z) - F_{p_0}(z)\big| &\le \sqrt{n}\sum_{x=0}^{z\wedge S}\big|\widehat{p}_n(x) - p_0(x)\big|\\ &\le (S+1)\sqrt{n}\|\widehat{p}_n - p_0\|_\infty = O_p(1).\end{aligned}$$

Also, the definition of $H_{\widehat{p}_n}$ and H_{p_0} and Proposition 2.6 imply that for all $z \in \mathbb{N}$

$$\begin{aligned}\sqrt{n}\big|H_{\widehat{p}_n}(z) - H_{p_0}(z)\big| &\le \sqrt{n}\sum_{x=0}^{z\wedge(S+1)-1}\big|F_{\widehat{p}_n}(x) - F_{p_0}(x)\big|\\ &\le (S+1)^2\sqrt{n}\|\widehat{p}_n - p_0\|_\infty = O_p(1)\end{aligned}$$

and the result follows. □

Proof of Proposition 2.6. First, note that the maximal point of the support of the empirical p.m.f. p_n is $X_{(n)} = \max_{1\le i\le n} X_i$, and that with probability one, $X_{(n)} = S$ provided that n is sufficiently large. Now, by Theorem 1 of [8] we know that $\widehat{p}_n$ admits a finite support whose maximal point $\widehat{s}_n \ge X_{(n)}$. Therefore, with probability one we have $\widehat{s}_n \ge S$ for n large enough. We show now by contradiction that with probability one there exists n^* such that if $n \ge n^*$ then $\widehat{s}_n \in \{S, S + 1\}$. Suppose that $\widehat{s}_n \ge S + 2$. By Proposition 1 of [8] we know that with probability one there exists n_0 such that for $n \ge n_0$, $\widehat{p}_n$ has to be linear on the set $\{S - 1, S, S + 1, \ldots, \widehat{s}_n\}$. But $S + 1$ is a

knot of p_0 which implies by Proposition 2.4 above that with probability one there exits $n^* \geq n_0$ such that for $n \geq n^*$, $S+1$ is also a knot of $\widehat{p}_n$. This yields a contradiction. □

Proof of Theorem 3.1. First, note that $\mathcal{C}(\mathcal{K})$ is a non-empty closed convex cone of $\mathbb{R}^{S+2}$, so there exists a unique minimizer of Φ over $\mathcal{C}(\mathcal{K})$.

Now suppose that $\widehat{g}$ is the minimizer of Φ over $\mathcal{C}(\mathcal{K})$. Let $x \in \{1, \ldots, S+2\}$. Then, for $\varepsilon > 0$ the function $k \mapsto \widehat{g}(k) + \varepsilon T_x(k)$, where $T_x(k) = (x-k)_+$, is clearly in $\mathcal{C}(\mathcal{K})$. Hence,

$$\begin{aligned} 0 &\leq \lim_{\varepsilon \searrow 0} \frac{\Phi(\widehat{g} + \varepsilon T_x) - \Phi(\widehat{g})}{\varepsilon} \\ &= \sum_{k=0}^{S+1} \widehat{g}(k) T_x(k) - \sum_{k=0}^{S+1} \mathbb{W}(k) T_x(k). \end{aligned}$$

Recall that $\widehat{\mathbb{G}}$ is the function defined by (3.6) for $k \in \{0, \ldots, S+1\}$, and $\mathbb{W}$ is given in (3.1). Setting $\widehat{\mathbb{G}}(-1) = 0$ for notational convenience, and noting that $F_{p_0}(-1) = 0$ implies that with probability one, $\mathbb{U}(F_{p_0}(-1)) = 0$, we can rewrite the last inequality as

$$\begin{aligned} 0 &\leq \sum_{k=0}^{S+1} \big\{\widehat{\mathbb{G}}(k) - \widehat{\mathbb{G}}(k-1)\big\}(x-k)_+ - \sum_{k=0}^{S+1} \big\{\mathbb{U}\big(F_{p_0}(k)\big) - \mathbb{U}\big(F_{p_0}(k-1)\big)\big\}(x-k)_+ \\ &= \sum_{k=0}^{S} \widehat{\mathbb{G}}(k)\big((x-k)_+ - \big(x-(k+1)\big)_+\big) + \widehat{\mathbb{G}}(S+1)(x-S-1)_+ \\ &\quad - \sum_{k=0}^{S} \mathbb{U}\big(F_{p_0}(k)\big)\big((x-k)_+ - \big(x-(k+1)\big)_+\big) - \mathbb{U}\big(F_{p_0}(S+1)\big)(x-S-1)_+ \\ &= \sum_{k=0}^{x-1} \widehat{\mathbb{G}}(k) - \sum_{k=0}^{x-1} \mathbb{U}\big(F_{p_0}(k)\big), \end{aligned}$$

where the last inequality follows from the fact that

$$(x-k)_+ - (x-k-1)_+ = \begin{cases} 1, & \text{if } k \leq x-1, \\ 0, & \text{if } k \geq x, \end{cases}$$

together with the fact that $(x-S-1)_+ = 0$ for $x \in \{1, \ldots, S+1\}$, and $(x-S-1)_+ = 1$ for $x = S+2$. Thus with probability one,

$$\widehat{\mathbb{H}}(x) = \sum_{k=0}^{x-1} \widehat{\mathbb{G}}(k) \geq \sum_{k=0}^{x-1} \mathbb{U}\big(F_{p_0}(k)\big) = \mathbb{H}(x)$$

for all $x \in \{1, \ldots, S+2\}$. Note that at $x = 0$, equality of $\widehat{\mathbb{H}}$ and $\mathbb{H}$ is guaranteed by the chosen convention. Proof of equality in case $\Delta\widehat{g}(x) > 0$ uses the fact that the perturbation function T_x

satisfies that $\widehat{g}+\varepsilon T_x$ is in $\mathcal{C}(\mathcal{K})$ for $|\varepsilon|$ small enough yielding

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\big(\Phi(\widehat{g}+\varepsilon T_x)-\Phi(\widehat{g})\big)=0.$$

We also have equality of $\widehat{\mathbb{H}}$ and $\mathbb{H}$ at the points in $\mathcal{K}\cup\{S+1,S+2\}$ since at these points there is no constraint.

Therefore, if $\widehat{g}$ is the minimizer of Φ, we have shown that the process $\widehat{\mathbb{H}}$ defined on $\{0,\dots,S+2\}$ as in (3.4) satisfies (3.5). Conversely, consider $\widehat{g}\in\mathcal{C}(\mathcal{K})$ such that the process $\widehat{\mathbb{H}}$ in (3.4) satisfies (3.5). Let $g\in\mathcal{C}(\mathcal{K})$. We will show now that $\Phi(g)\geq\Phi(\widehat{g})$. We have

$$\begin{aligned}\Phi(g)-\Phi(\widehat{g})&=\frac{1}{2}\sum_{k=0}^{S+1}\big(g(k)-\widehat{g}(k)\big)^2+\sum_{k=0}^{S+1}\big(g(k)-\widehat{g}(k)\big)\big(\widehat{g}(k)-\mathbb{W}(k)\big)\\&\geq\sum_{k=0}^{S+1}\big(g(k)-\widehat{g}(k)\big)\big(\widehat{g}(k)-\mathbb{W}(k)\big)\\&=\sum_{k=0}^{S+1}\big(g(k)-\widehat{g}(k)\big)\big(\widehat{\mathbb{D}}(k)-\widehat{\mathbb{D}}(k-1)\big),\end{aligned}$$

where $\widehat{\mathbb{D}}(k)=\widehat{\mathbb{G}}(k)-\mathbb{U}(F_{p_0}(k))$. Now, similar as above, for all $x\in\{1,\dots,S+2\}$ we have

$$\begin{aligned}\sum_{k=0}^{S+1}\big(\widehat{\mathbb{D}}(k)-\widehat{\mathbb{D}}(k-1)\big)(x-k)_+&=\sum_{k=0}^{x-1}\widehat{\mathbb{D}}(k)\\&=\widehat{\mathbb{H}}(x)-\mathbb{H}(x)\geq 0,\end{aligned}\tag{5.4}$$

with equality if $\Delta\widehat{g}(x)>0$, or a point in $\mathcal{K}\cup\{0,S+1,S+2\}$. To conclude, we will use the fact that an arbitrary element $g\in\mathcal{C}(\mathcal{K})$ can be written as

$$g(k)=\alpha+\sum_{j=1}^{m+1}c_j(s_j-k)_+ +\sum_{j=1}^{m+1}\sum_{i=1}^{J_j}c_{j,i}(z_{j,i}-k)_+\tag{5.5}$$

for all $k=0,\dots,S+1$, with $s_1<\cdots<s_m$ the interior knots of p_0, $s_{m+1}=S+1$, $z_{j,1},\dots,z_{j,J_j}$ the knots of g in $\{s_{j-1}+1,\dots,s_j-1\}$ for $j=1,\dots,m+1$, and where $\alpha,c_1,c_2,\dots,c_{m+1}$ are real numbers, and $c_{j,i}>0$ for $j=1,\dots,m+1$ and $i=1,\dots,J_j$. This comes from the fact that any finite convex sequence $p=\{p(0),\dots,p(K)\}$ for some $K>0$, admits the (spline) representation

$$p(k)=a+\gamma_1(s_1-k)_+ +\cdots+\gamma_p(s_p-k)_+ +\gamma_{p+1}(K-k)_+,\tag{5.6}$$

where a is a real number, $\gamma_i>0$ and $0<s_1<\cdots<s_p<K$ are the interior knots of p. Using the spline representation in (5.5) together with (5.4), where we recall that we have an equality

for $x = s_1, \dots, s_{m+1}$, it follows that

$$\begin{aligned}
&\sum_{k=0}^{S+1} \big(\widehat{\mathbb{D}}(k) - \widehat{\mathbb{D}}(k-1)\big) g(k) \\
&\quad = \alpha \widehat{\mathbb{D}}(S+1) + \sum_{j=1}^{m+1} \sum_{i=1}^{J_j} c_{j,i} \sum_{k=0}^{S+1} \big(\widehat{\mathbb{D}}(k) - \widehat{\mathbb{D}}(k-1)\big)(z_{j,i} - k)_+ \\
&\quad \geq \alpha \widehat{\mathbb{D}}(S+1),
\end{aligned}$$

where in the last inequality we used the fact that $c_{j,i} \geq 0$ for all j, i. Now, the boundary conditions $\widehat{\mathbb{H}}(S+1) = \mathbb{H}(S+1)$ and $\widehat{\mathbb{H}}(S+2) = \mathbb{H}(S+2)$ in (3.5) imply that

$$\begin{aligned}
\widehat{\mathbb{D}}(S+1) &= \widehat{G}(S+1) - \mathbb{U}\big(F_{p_0}(S+1)\big) \\
&= \widehat{\mathbb{H}}(S+2) - \widehat{\mathbb{H}}(S+1) - \mathbb{U}\big(F_{p_0}(S+1)\big) \\
&= \mathbb{H}(S+2) - \mathbb{H}(S+1) - \mathbb{U}\big(F_{p_0}(S+1)\big) = 0.
\end{aligned}$$

We arrive at

$$\sum_{k=0}^{S+1} \big(\widehat{\mathbb{D}}(k) - \widehat{\mathbb{D}}(k-1)\big) g(k) \geq 0$$

and similarly, since we have an equality in (5.4) if $\Delta \widehat{g}(x) > 0$,

$$\sum_{k=0}^{S+1} \big(\widehat{\mathbb{D}}(k) - \widehat{\mathbb{D}}(k-1)\big) \widehat{g}(k) = 0.$$

It follows that $\Phi(g) \geq \Phi(\widehat{g})$ and that $\widehat{g}$ is the minimizer of Φ. □

Proof of Theorem 3.2. For $z \in \mathbb{N}$ define

$$\begin{aligned}
\mathbb{Y}_n(z) &=: \sum_{k=0}^{z-1} \sqrt{n}\big(F_{p_n}(k) - F_{p_0}(k)\big) = \sum_{k=0}^{z-1} \mathbb{U}_n\big(F_{p_0}(k)\big), \\
\widehat{Y}_n(z) &=: \sum_{k=0}^{z-1} \sqrt{n}\big(F_{\widehat{p}_n}(k) - F_{p_0}(k)\big),
\end{aligned}$$

and recall that $\mathbb{H}$ is given by (3.3) with $\mathbb{Y}_n(0) = \widehat{Y}_n(0) = \mathbb{H}(0) = 0$. It follows from the characterization of $\widehat{p}_n$ in (2.6) that

$$\widehat{Y}_n(x) \begin{cases} \geq \mathbb{Y}_n(x), & \text{for } x \in \{0, \dots, S+2\}, \\ = \mathbb{Y}_n(x), & \text{if } x \in \{0, \dots, S+2\} \text{ is a knot of } \widehat{p}_n. \end{cases} \tag{5.7}$$

By standard results on weak convergence of empirical processes, we have the joint convergence

$$\mathbb{Y}_n(x) \underset{d}{\to} \mathbb{H}(x), \qquad \text{for } x \in \{0, \ldots, S+2\}.$$

Now, Theorem 2.5 implies that there exists a subsequence $\{\widehat{Y}_{n'}\}_{n'}$ which weakly converges to some $\mathbb{L}$ on $\{0, \ldots, S+2\}$. In what follows, we will use the Skorokhod representation to assume that convergences of $\{\widehat{Y}_{n'}\}_{n'}$ and $\{\mathbb{Y}_{n'}\}_{n'}$ to their respective limits happen almost surely. The goal now is to show that $\mathbb{L}$ and $\widehat{\mathbb{H}}$, defined in Theorem 3.1 above, are equal with probability one. Let us define

$$\widetilde{g}(x) := \mathbb{L}(x+1) + \mathbb{L}(x-1) - 2\mathbb{L}(x)$$

for $x \in \{0, \ldots, S+1\}$ with $\mathbb{L}(-1) = 0$. Note that we have that

$$\mathbb{L}(S+2) = \mathbb{L}(S+1) = \mathbb{H}(S+1) = \mathbb{H}(S+2). \tag{5.8}$$

Indeed, it follows from Proposition 2.6 that $\widehat{p}_n$ is a p.m.f. whose support is included in $\{0, \ldots, S+1\}$ with probability one, for sufficiently large n. This implies that $F_{\widehat{p}_n}(S+1) = F_{p_0}(S+1) = 1$ and therefore, $\widehat{\mathbb{Y}}_n(S+2) = \widehat{\mathbb{Y}}_n(S+1)$. Similarly, $\mathbb{Y}_n(S+2) = \mathbb{Y}_n(S+1)$ with probability one. Now, $S+1$ is a knot of p_0 so it follows from Proposition 2.4 that $S+1$ is also a knot of $\widehat{p}_n$ with probability one, for n sufficiently large. Hence, by (5.7) we have $\widehat{\mathbb{Y}}_n(S+1) = \mathbb{Y}_n(S+1)$. We conclude that with probability one and n large enough,

$$\widehat{\mathbb{Y}}_n(S+2) = \widehat{\mathbb{Y}}_n(S+1) = \mathbb{Y}_n(S+1) = \mathbb{Y}_n(S+2),$$

and the claim follows by passing $n' \to \infty$. Hence, with probability one we have that

$$\mathbb{L}(x) \begin{cases} \geq \mathbb{H}(x), & \text{for } x \in \{0, \ldots, S+2\}, \\ = \mathbb{H}(x), & \text{if } x \in \mathcal{K} \cup \{0, S+1, S+2\} \text{ or } \Delta\widetilde{g}(x) > 0. \end{cases}$$

Equality of $\mathbb{L}$ and $\mathbb{H}$ at points in $\mathcal{K} \cup \{0, S+1, S+2\}$ in the last assertion follows from the chosen convention at 0, Proposition 2.4 together with the equalities $\widehat{Y}_{n'}(x) = \mathbb{Y}_{n'}(x)$ with probability one and n' large enough for $x \in \mathcal{K}$, and the identity in (5.8) for $S+1$ and $S+2$. Equality of $\mathbb{L}$ and $\mathbb{H}$ at points $x \in \{1, \ldots, S\} \setminus \mathcal{K}$ with $\Delta\widetilde{g}(x) > 0$ follows from the fact that x is a knot of $\widehat{p}_{n'}$ in $\{s+1, \ldots, s'-1\}$, with s and s' being two successive knots of p_0, if and only if it is a knot of $\sqrt{n'}(\widehat{p}_{n'} - p_0)$ using linearity of p_0 between two successive knots. Thus, if x is a knot of $\widetilde{g}$, then as soon as n' is large enough x is also a knot of

$$z \mapsto \widehat{Y}_{n'}(z+1) + \widehat{Y}_{n'}(z-1) - 2\widehat{Y}_{n'}(z) = \sqrt{n'}\big(\widehat{p}_{n'}(z) - p_0(z)\big).$$

This in turn implies that $\widehat{Y}_{n'}(x) = \mathbb{Y}_{n'}(x)$ implying after passing to the limit that $\mathbb{L}(x) = \mathbb{H}(x)$ almost surely. Furthermore, $\widetilde{g}$ is clearly in $\mathcal{C}(\mathcal{K})$.

Therefore, it follows from Theorem 3.1 that $\widetilde{g}$ must be equal to the minimizer of Φ defined in (3.2). Thus, there exists a version of $\widehat{g}$ such that

$$\widehat{g}(x) = \mathbb{L}(x+1) + \mathbb{L}(x-1) - 2\mathbb{L}(x) = \widehat{\mathbb{H}}(x+1) + \widehat{\mathbb{H}}(x-1) - 2\widehat{\mathbb{H}}(x)$$

for $x \in \{0, \ldots, S+1\}$. Put $\Delta = \widehat{\mathbb{H}} - \mathbb{L}$. Then, for $x \in \{0, \ldots, S+1\}$, we have that $\Delta(x+1) = 2\Delta(x) - \Delta(x-1)$. But $\Delta(-1) = \Delta(0) = 0$ since $\mathbb{L}(-1) = \widehat{\mathbb{H}}(-1) = 0$ and $\mathbb{L}(0) = \widehat{\mathbb{H}}(0) = 0$. We conclude by induction that $\Delta = 0$, that is $\widehat{\mathbb{H}} = \mathbb{L}$, on $\{0, \ldots, S+2\}$. Now, from an arbitrary subsequence n' we can extract a further subsequence n'' along which $\widehat{\mathbb{Y}}_{n''}$ and $\mathbb{Y}_{n''}$ weakly converge jointly to $\widehat{\mathbb{H}}$ and $\mathbb{H}$. Since the limit is the same for any such subsequence, we conclude that $\widehat{\mathbb{Y}}_n$ and $\mathbb{Y}_n$ weakly converge jointly to $\widehat{\mathbb{H}}$ and $\mathbb{H}$ on $\{0, \ldots, S+2\}$. This in turn implies that the following convergences

$$\sqrt{n}(H_{\widehat{p}_n} - H_{p_0}) \Rightarrow \widehat{\mathbb{H}}, \qquad \sqrt{n}(F_{\widehat{p}_n} - F_{p_0}) \Rightarrow \widehat{\mathbb{G}}$$

and $\sqrt{n}(\widehat{p}_n - p_0) \Rightarrow \widehat{g}$ occur jointly, and the proof is complete. □

Proof of Theorem 3.3. Let $\mathcal{K}_n$ be the set of all points $x \in \{1, \ldots, S\}$ that satisfy $\Delta\widehat{p}_n(x) > v_n$. Let $\varepsilon = \min_{x \in \mathcal{K}} \Delta p_0(x)$. Since $\mathcal{K}$ is the set of all interior knots of p_0, we have $\varepsilon > 0$ whenever $\mathcal{K} \neq \varnothing$. We will now prove that $\lim_{n\to\infty} P(\mathcal{K}_n \neq \mathcal{K}) = 0$. To this end, note that

$$\begin{aligned} P(\mathcal{K}_n \neq \mathcal{K}) \leq{}& P(\text{there exists } x \in \mathcal{K} \text{ such that } x \notin \mathcal{K}_n) \\ &+ P(\text{there exists } x \in \mathcal{K}_n \text{ such that } x \notin \mathcal{K}), \end{aligned} \tag{5.9}$$

whence it suffices to show that both probabilities on the right-hand side tend to zero.

Consider the first probability. If there exists $x \in \mathcal{K}$ such that $x \notin \mathcal{K}_n$, then $\mathcal{K}$ is not empty, and it follows from the definition of ε and $\mathcal{K}_n$ that $\Delta\widehat{p}_n(x) \leq v_n$ and $\Delta p_0(x) \geq \varepsilon$. Since $\lim_{n\to\infty} v_n = 0$, this means that

$$\Delta\widehat{p}_n(x) - \Delta p_0(x) \leq -\varepsilon/2$$

for n sufficiently large. Hence, the first probability is bounded from above by

$$P\big(\text{there exists } x \in \{1, \ldots, S\} \text{ such that } \Delta\widehat{p}_n(x) - \Delta p_0(x) \leq -\varepsilon/2\big) \tag{5.10}$$

for n sufficiently large. However, it follows from Theorem 3.2 that $\Delta\widehat{p}_n(x) - \Delta p_0(x) = O_p(n^{-1/2})$ for all $x \in \{1, \ldots, S\}$ and therefore, the probability in (5.10) converges to zero, implying that the first probability on the right-hand side of (5.9) converges to zero as $n \to \infty$.

Next, consider the second probability on the right side of (5.9). If there exists $x \in \mathcal{K}_n$ such that $x \notin \mathcal{K}$, then $\Delta\widehat{p}_n(x) > v_n$ and $\Delta p_0(x) = 0$. Since $\lim_{n\to\infty} \sqrt{n}v_n = \infty$, we can find $N \in \mathbb{N}$ such that for a given $A > 0$ we have that $\sqrt{n}v_n > A$ for all $n \geq N$. This in turn implies that

$$\sqrt{n}\big(\Delta\widehat{p}_n(x) - \Delta p_0(x)\big) > A$$

for all $n \geq N$. Fix $\eta > 0$. Applying again Theorem 3.2, we can choose A sufficiently large so that

$$\begin{aligned} &P(\text{there exists } x \in \mathcal{K}_n \text{ such that } x \notin \mathcal{K}) \\ &\quad \leq P\big(\text{there exists } x \in \{1, \ldots, S\} \text{ such that } \sqrt{n}\big(\Delta\widehat{p}_n(x) - \Delta p_0(x)\big) > A\big) < \eta \end{aligned}$$

for all $n \geq N$. Since η was arbitrary, this implies that the second probability on the right hand side of (5.9) converges to 0 and that $\lim_{n\to\infty} P(\mathcal{K}_n \neq \mathcal{K}) = 0$.

On the other hand, $S_n = S$ with probability converging to 1. Hence, we can assume without loss of generality that $S_n = S$ and $\mathcal{K}_n = \mathcal{K}$. In this case, we also have $\mathcal{C}_n = \mathcal{C}(\mathcal{K})$ and $\widehat{g}_n$ is precisely the minimizer of

$$\frac{1}{2}\sum_{k=0}^{S+1}\big(g(k) - g_n(k)\big)^2$$

over $\mathcal{C}(\mathcal{K})$, which means that $\widehat{g}_n$ is the convex projection of g_n on the non-empty closed convex set $\mathcal{C}(\mathcal{K})$ in $\mathbb{R}^{S+2}$.

Now, assume that $X_1, \ldots, X_n$ are the first n terms of a sequence $(X_i)_{i\in\mathbb{N}}$ of i.i.d. random variables, and let G be a standard Gaussian vector with dimension $S+2$, which is independent of $(X_i)_{i\in\mathbb{N}}$. Conditionally on $(X_1, \ldots, X_n)$, the random vector $(g_n(0), \ldots, g_n(S+1))$ has the same distribution as

$$\Gamma_n^{1/2} G.$$

Moreover, with Γ being the $(S+2)\times(S+2)$ matrix with component $(i+1, j+1)$ equal to $p_0(i)(1-p_0(i))$ for all $i = j$ and $-p_0(i)p_0(j)$ for all $i \neq j$, with $i, j = 0, \ldots, S+1$, the random vector $(\mathbb{W}(0), \ldots, \mathbb{W}(S+1))$ has the same distribution as

$$\Gamma^{1/2} G.$$

In the sequel, we assume without loss of generality that $(g_n(0), \ldots, g_n(S+1)) = \Gamma_n^{1/2} G$ and $(\mathbb{W}(0), \ldots, \mathbb{W}(S+1)) = \Gamma^{1/2} G$. Thus in particular, $\mathbb{W}$ and g_n are defined on the same probability space. From what precedes, $\widehat{g}_n$ is the convex projection of g_n on the non-empty closed convex set $\mathcal{C}(\mathcal{K})$ in $\mathbb{R}^{S+2}$ whereas $\widehat{g}$ was defined to be the convex projection of $\mathbb{W}$ on the same set $\mathcal{C}(\mathcal{K})$ in $\mathbb{R}^{S+2}$. Using that the metric projection is Lipschitz, we conclude that

$$\sum_{k=0}^{S+1}\big(\widehat{g}(k) - \widehat{g}_n(k)\big)^2 \le \sum_{k=0}^{S+1}\big(\mathbb{W}(k) - g_n(k)\big)^2.$$

From the Cauchy–Schwarz inequality, it follows that

$$\sum_{k=0}^{S+1}\big(\widehat{g}(k) - \widehat{g}_n(k)\big)^2 \le \sum_{i=1}^{S+2}\sum_{j=1}^{S+2}\big(\Gamma_{n,i,j}^{1/2} - \Gamma_{i,j}^{1/2}\big)^2 \times \sum_{k=1}^{S+2} G_k^2, \tag{5.11}$$

where $\Gamma_{n,i,j}$, $\Gamma_{i,j}$ and G_k, respectively denote the component (i, j) of Γ_n, the component (i, j) of Γ, and the component k of G. But p_n almost surely converges to p_0 on $\{0, \ldots, S+2\}$, so Γ_n converges to Γ conditionally on $(X_i)_{i\in\mathbb{N}}$, with probability one. This means that the right-hand side in (5.11) converges to zero conditionally on $(X_i)_{i\in\mathbb{N}}$, with probability one. As this implies that the left hand side converges to zero as well, this proves that $\widehat{g}_n$ converges in distribution to $\widehat{g}$ conditionally on $(X_i)_{i\in\mathbb{N}}$, which completes the proof of Theorem 3.3. □

Proof of Theorem 3.6. Existence and uniqueness of the minimizer both follow from the projection theorem on closed convex cones in $\mathbb{R}^{s+1}$. With similar arguments as for the proof of

Theorem 3.1 and using that $s \in \mathcal{K}$, it can be shown that an arbitrary element $\widehat{g}^{\le s} \in \mathcal{C}^{\le s}(\mathcal{K})$ is the minimizer if and only if the process $\widehat{\mathbb{H}}^{\le s}$ defined on $\{0, \ldots, s+1\}$ by

$$\widehat{\mathbb{H}}^{\le s}(x) = \sum_{k=0}^{x-1} \sum_{j=0}^{k} \widehat{g}^{\le s}(j),$$

if $x \in \{1, \ldots, s+1\}$ and $\widehat{\mathbb{H}}^{\le s}(0) = 0$ satisfies

$$\widehat{\mathbb{H}}^{\le s}(x) \ge \mathbb{H}(x)$$

for all $x \in \{0, \ldots, s+1\}$, with an equality if $x \in \mathcal{K} \cup \{0, s+1\}$ or $x \in \{1, \ldots, s-1\}$ satisfies $\Delta\widehat{g}^{\le s}(x) > 0$. Consider the restriction $\widehat{g}^{\le s} = (\widehat{g}(0), \ldots, \widehat{g}(s))$ of $\widehat{g}$ to $\{0, \ldots, s\}$. A point $x \in \{1, \ldots, s-1\}$ is a knot of $\widehat{g}^{\le s}$ if, and only if, it is a knot of $\widehat{g}$. Therefore, it immediately follows from the characterization of $\widehat{g}$ given in Theorem 3.1 that the minimizer of $\Phi^{\le s}$ over $\mathcal{C}^{\le s}$ is equal to $\widehat{g}^{\le s}$ if, and only if, $\widehat{\mathbb{H}}(s+1) = \mathbb{H}(s+1)$. Since we already have $\widehat{\mathbb{H}}(s) = \mathbb{H}(s)$, this is equivalent to

$$\widehat{\mathbb{H}}(s+1) - \widehat{\mathbb{H}}(s) = \mathbb{H}(s+1) - \mathbb{H}(s),$$

that is (3.9).

To prove the last assertion, note that from Theorem 3.2, it follows that $\sqrt{n}(\widehat{p}_n(0) - p_0(0), \ldots, \widehat{p}_n(s) - p_0(s))$ converges in distribution to $(\widehat{g}(0), \ldots, \widehat{g}(s))$ as $n \to \infty$. Thus, it converges in distribution to the minimizer of $\Phi^{\le s}$ over $\mathcal{C}^{\le s}(\mathcal{K})$ if, and only if, (3.9) holds true with probability one. On the other hand, from Theorem 3.2, one also has

$$\sqrt{n}\begin{pmatrix} F_{\widehat{p}_n} - F_{p_0} \\ F_{p_n} - F_{p_0} \end{pmatrix} \Rightarrow \begin{pmatrix} \widehat{\mathbb{G}} \\ \mathbb{U}_n(F_{p_0}) \end{pmatrix} \tag{5.12}$$

as $n \to \infty$. Therefore, one can have (3.9) with probability one if, and only if,

$$\sqrt{n}\big(\big(F_{\widehat{p}_n}(s) - F_{p_0}(s)\big) - \big(F_{p_n}(s) - F_{p_0}(s)\big)\big)$$

converges in probability to zero as $n \to \infty$. This is equivalent to (3.10), so the proof of the theorem is complete. □

Proof of Theorem 3.7. Existence and uniqueness of the minimizer both follow from the projection theorem on closed convex cones in $\mathbb{R}^{S-s}$. With similar arguments as for the proof of Theorem 3.1, it can be shown that an arbitrary element $\widehat{g}^{\ge s} \in \mathcal{C}^{\ge s}(\mathcal{K})$ is the minimizer if and only if the process $\widehat{\mathbb{H}}^{\ge s}$ defined on $\{s, \ldots, S+2\}$ by

$$\widehat{\mathbb{H}}^{\ge s}(x) = \sum_{k=s}^{x-1} \sum_{j=s}^{k} \widehat{g}^{\ge s}(j),$$

if $x \in \{s+1, \ldots, S+2\}$ and $\widehat{\mathbb{H}}^{\ge s}(s) = 0$ satisfies $\widehat{\mathbb{H}}^{\ge s}(x) \ge \mathbb{H}^{\ge s}(x)$ for all $x \in \{s, \ldots, S+2\}$, with an equality if $x \in \mathcal{K} \cup \{0, S+1, S+2\}$ or $x \in \{1, \ldots, S\}$ satisfies $\Delta\widehat{g}^{\ge s}(x) > 0$, where

$\mathbb{H}^{\geq s}(s) = 0$ and

$$\mathbb{H}^{\geq s}(x) = \sum_{k=s}^{x-1} \sum_{j=s}^{k} \mathbb{W}(j)$$

for all $x \in \{s+1, \ldots, S+2\}$. The connection between $\mathbb{H}$ and $\mathbb{H}^{\geq s}$ is as follows. For all $x \in \{s, \ldots, S+2\}$ one has

$$\begin{aligned}\mathbb{H}(x) &= \sum_{k=0}^{s-1} \sum_{j=0}^{k} \mathbb{W}(j) + \sum_{k=s}^{x-1} \sum_{j=0}^{s-1} \mathbb{W}(j) + \mathbb{H}^{\geq s}(x) \\ &= \mathbb{H}(s) + (x-s)\big(\mathbb{H}(s) - \mathbb{H}(s-1)\big) + \mathbb{H}^{\geq s}(x).\end{aligned}$$

Similarly,

$$\widehat{\mathbb{H}}(x) = \widehat{\mathbb{H}}(s) + (x-s)\big(\widehat{\mathbb{H}}(s) - \widehat{\mathbb{H}}(s-1)\big) + \widehat{\mathbb{H}}^{\geq s}(x)$$

for all $x \in \{s, \ldots, S+2\}$, so that

$$\widehat{\mathbb{H}}^{\geq s}(x) - \mathbb{H}^{\geq s}(x) = \widehat{\mathbb{H}}(x) - \mathbb{H}(x) + (x-s)\big(\widehat{\mathbb{H}}(s-1) - \mathbb{H}(s-1)\big), \tag{5.13}$$

using that $\widehat{\mathbb{H}}(s) = \mathbb{H}(s)$.

Consider $\widehat{g}^{\geq s} = (\widehat{g}(s), \ldots, \widehat{g}(S+1))$. A point $x \in \{s+1, \ldots, S\}$ is a knot of $\widehat{g}^{\geq s}$ if, and only if, it is a knot of $\widehat{g}$. Therefore, it immediately follows from the characterization of $\widehat{g}$ given in Theorem 3.1, together with (5.13), that the minimizer of $\Phi^{\geq s}$ over $\mathcal{C}^{\geq s}(\mathcal{K})$ is equal to $(\widehat{g}(s), \ldots, \widehat{g}(S+1))$ if, and only if, $\widehat{\mathbb{H}}(s-1) = \mathbb{H}(s-1)$. Since we already have $\widehat{\mathbb{H}}(s) = \mathbb{H}(s)$, this is equivalent to

$$\widehat{\mathbb{H}}(s) - \widehat{\mathbb{H}}(s-1) = \mathbb{H}(s) - \mathbb{H}(s-1),$$

that is (3.11).

To prove the last assertion, note that from Theorem 3.2, it follows that $\sqrt{n}(\widehat{p}_n(s) - p_0(s), \ldots, \widehat{p}_n(S+1) - p_0(S+1))$ converges in distribution to $(\widehat{g}(s), \ldots, \widehat{g}(S+1))$ as $n \to \infty$. Thus, it converges in distribution to the minimizer of $\Phi^{\geq s}$ over $\mathcal{C}^{\geq s}(\mathcal{K})$ if, and only if, (3.11) holds true with probability one. According to (5.12), this happens if, and only if,

$$\sqrt{n}\Big(\big(F_{\widehat{p}_n}(s-1) - F_{p_0}(s-1)\big) - \big(F_{p_n}(s-1) - F_{p_0}(s-1)\big)\Big)$$

converges in probability to zero as $n \to \infty$. This is equivalent to (3.12), so the proof of the theorem is complete. □

Proof of Lemma 4.2. We have $\Delta p(k) \geq 0$ for all $k < S$ by convexity of the function $k \mapsto q^k$. We also have $\Delta p(k) \geq 0$ for all integers $k > S+1$ since for those k, $\Delta p(k) = 0$. Now, $\Delta p(S+1) = p(S) \geq 0$, and

$$\Delta p(S) = \frac{q^{S-1}(1-q)}{1-q^{S+1}}(1-2q),$$

which is ≥ 0 if, and only if, $q \leq 1/2$. We conclude that $\Delta p(k) \geq 0$ for all $k \in \mathbb{N}$ if, and only if, $q \leq 1/2$. □

Acknowledgements

The authors are very grateful to Sylvie Huet for many helpful comments, stimulating discussions on the subject, and also for a careful reading of the paper.

References

[1] Balabdaoui, F. and Durot, C. (2015). Marshall lemma in discrete convex estimation. *Statist. Probab. Lett.* **99** 143–148. MR3321508

[2] Balabdaoui, F., Jankowski, H., Rufibach, K. and Pavlides, M. (2013). Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 769–790. MR3091658

[3] Balabdaoui, F., Rufibach, K. and Wellner, J.A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331. MR2509075

[4] 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*. New York: Wiley. MR0326887

[5] Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798

[6] Dümbgen, L. and Rufibach, K. (2010). Logcondens: Computations related to univariate log-concave density estimation. *J. Stat. Softw.* **39** Article ID 6.

[7] Dümbgen, L., Rufibach, K. and Wellner, J.A. (2007). Marshall's lemma for convex density estimation. In *Asymptotics: Particles, Processes and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 101–107. Beachwood, OH: IMS. MR2459933

[8] 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. MR3079603

[9] Durot, C., Huet, S., Koladjo, F. and Robin, S. (2015). Nonparametric species richness estimation under convexity constraint. *Environmetrics* **26** 502–513.

[10] Durot, C., Kulikov, V.N. and Lopuhaä, H.P. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *Ann. Statist.* **40** 1578–1608. MR3015036

[11] Dykstra, R.L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842. MR0727568

[12] Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

[13] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

[14] Groeneboom, P., Jongbloed, G. and Wellner, J.A. (2008). The support reduction algorithm for computing non-parametric function estimates in mixture models. *Scand. J. Statist.* **35** 385–399. MR2446726

[15] Holm, S. and Frisn, M. (1985). Nonparametric regression with simple curve characteristics. Research report 4, Dept. Statistics, Univ. Göteborg.

[16] Jankowski, H.K. and Wellner, J.A. (2009). Estimation of a discrete monotone distribution. *Electron. J. Stat.* **3** 1567–1605. MR2578839

[17] Kulikov, V.N. and Lopuhaä, H.P. (2005). Asymptotic normality of the L_k-error of the Grenander estimator. *Ann. Statist.* **33** 2228–2255. MR2211085

[18] Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. MR1105842

[19] Ng, P. and Maechler, M. (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Stat. Model.* **7** 315–328. MR2749902

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

Received April 2014 and revised March 2015