2017-06-09

Constrained maximum likelihood estimators for densities

Royset, Johannes O.; Wets, Roger J.-B.

J.O. Royset, R.J.-B. Wets, "Constrained maximum likelihood estimators for densities," arXic:1702.08109v3 [math.ST] 9 Jun 2017
https://hdl.handle.net/10945/56666

Constrained Maximum Likelihood Estimators for Densities

Johannes O. Royset

Operations Research Department
Naval Postgraduate School
joroyset@nps.edu

Roger J-B Wets

Department of Mathematics
University of California, Davis
rjbwets@ucdavis.edu

Abstract. We propose a framework for nonparametric maximum likelihood estimation of densities in situations where the sample is supplemented by information and assumptions about shape, support, continuity, slope, location of modes, density values, and other conditions that, individually or in combination, restrict the family of densities under consideration. We establish existence of estimators and their cluster points, strong consistency under mild assumptions, and robustness in the presence of model misspecification. The results are achieved by means of viewing densities as elements of spaces of semicontinuous functions with the hypo-distance metric. This metric emerges as natural and convenient when considering broad classes of side conditions. It also has the exceptional property that convergence of densities in this metric implies convergence of modes, near-modes, height of modes, and high-likelihood events. Thus, we automatically achieve strong consistency of a rich class of plug-in estimators for modes and related quantities. Relying on almost sure epi-convergence of criterion functions, we avoid the strong assumptions associated with uniform laws of large numbers and instead leverage a less demanding law, for which we provide a new proof. Specific examples illustrate the framework including an estimator simultaneously subject to bounds on density values and its (sub)gradients, restriction to concavity, penalization that encourages lower modes, and imprecise information about the expected value.

Keywords: nonparametric density estimation, maximum likelihood, shape-constrained estimation, consistency, variational approximations, epi-convergence, hypo-distance, epi-splines.

Date: June 12, 2017

1 Introduction

It is self-evident that statistical estimates should rely on all available information about the pertinent stochastic phenomenon. Therefore, observation-data needs to be supplemented by whatever is known, or reasonably reckoned, about its distribution: shape, bounds on moments, slope, modes, support, tail characteristics, proximity to a "prior" distribution and so on. When any combination of these properties is included in the formulation of an estimation problem, it mutates into a more complex one

and deriving consistency results, computing estimates, and even establishing existence of estimators and their cluster points become significantly more challenging. In this article, we deal with these issues in the context of nonparametric maximum likelihood estimation of density functions on $\mathbb{R}^d$.

It is already widely accepted that shape restrictions enrich modeling possibilities, regularize estimators, reduce the need for nuisance parameters, and improve the generalization error. They offer possibilities between inflexible parametric models and data-intensive ones from traditional nonparametric statistics. Maximum likelihood estimators constrained by shape restrictions and other side conditions have the appealing property that for any finite sample size they still satisfy the constraints and therefore tend to perform well in practically challenging situations with relatively small sample sizes. For the same reason they are also easily interpreted and explained as resulting in the "most likely convex density on support S with moments in set C," or whatever constraints are imposed in a particular application.

The ability to handle constraints enables statistical modeling that resembles, at a high level, those in modern machine learning where features and layers are engineered through an extensive process of cross-validation and experimentation, often based on domain knowledge and experience. Density estimators constructed in this paper can incorporate essentially any combination of constraints arising from a similar process. Our framework represents a departure from traditional shape constrained density estimators with mostly one, exceptionally two, side conditions to those where constraints take the primary role.

The estimators developed here fall within the broad class of M-estimators for which it is already well understood that a central challenge is to establish a continuity property of the argmin-mapping under some appropriate topology on the space of criterion functions. Classical results typically require the argmin to be a singleton for the criterion function corresponding to the actual (true) probability measure; see for example Theorem 3.2.2 and Corollary 3.2.3 in [57]. Extensions include results on rates of convergence that rely on the bracketing numbers of the families of functions (densities) from which the estimators are selected as well as the speed with which one can approach the actual function (density) using elements from these families; see Theorem 3.4.4 in [57] and Theorem 8.12 in [59] for maximum likelihood estimators of densities in terms of the Hellinger distance.

Serious challenges arise when families of densities under consideration are restricted by nontrivial constraints. First, the argmin of an empirical criterion function (using an empirical measure) may be empty and thus an estimator fails to exist. This is known to take place even in relatively simple cases such as under unimodality constraints [7] and emerges as a significant issue in the present context with much more general constraints that may even change with the sample size. Traditionally, the existence of an M-estimator is established by ad hoc means in specific cases. In this paper, we provide a systematic approach and illustrate it for constraints related to convexity, concavity, log-concavity, s-concavity, monotone transformations, monotonicity, Lipschitz continuity, pointwise upper and lower bounds, location of modes, height at modes, values of moments, size of subgradients, approximate evaluation of the integral of densities, restrictions to splines, restrictions to multivariate totally positive densities of order two, and *any combination* of the above. To the best of our knowledge, no prior study has established existence of maximum likelihood estimators for such a variety of constraints.

A second challenge is that the argmin of the actual criterion function (based on the actual measure) could be empty, not a singleton, or noncompact. In the density estimation context the situation is most troublesome when the imposed constraints *exclude* the actual density, i.e., we are faced with model misspecification, in which case the problem of likelihood optimization becomes that of minimizing the Kullback-Leibler divergence to the actual density. Although it has been addressed in a few specific cases such as those under log-concavity constraints [10], it is generally nontrivial to establish that this minimization problem in the presence of constraints has a nonempty compact solution set in the appropriate topology. Thus, even if estimators exist, they may not have cluster points and consistency is in jeopardy. Model misspecification is highly plausible in the present context where side conditions are envisioned being added aggressively based on experience and intuition. We guarantee the existence of cluster points and that all such points indeed minimize the Kullback-Leibler divergence to the actual density almost surely under mild assumptions. Thus, our estimators exhibit a certain kind of robustness under model misspecification. Concrete results illustrate the approach for the wide variety of constraints and their combinations listed in the previous paragraph.

It is apparent that argmin-mappings in our general setting with nontrivial constraints need to be examined using a topology on families of criterion functions that ensures convergence of *sets* of optimal solutions even in the presence of constraints that could depend on the sample size. This leads us to the topology generated by *epi-convergence* that has emerged as the principal path for establishing continuity of set-valued argmin-mappings in optimization theory [47, Chapter 7]. With rare exceptions (see for example [16] in the density setting and [20, 61] in the regression context), this is a path to estimator consistency not utilized in the statistics literature where the preference is to pass through uniform convergence of the criterion functions (see for example Theorem 3.2.2 in [57]), which is generally more demanding as we see in Section 3.2 when we present a law of large numbers for epi-convergence. In particular, our consideration of epi-convergence for the criterion functions defined on the selected families of densities described below is novel.

As in all infinite-dimensional analysis, the choice of metric on the family of densities under consideration may have profound impact on the usefulness of the results and ease with which they are achieved. The Hellinger distance is in some sense natural on spaces of densities (see for example Theorem 3.4.4 in [57]), but the introduction of increasingly complex constraints results in significant difficulties due, at least in part, to the potential lack of compactness of the family of permissible densities. In this paper, we examine for the first time families of *semicontinuous densities equipped with the hypo-distance metric*, which naturally emerge as *compact*. In fact, *every* bounded family of nonnegative extended real-valued upper semicontinuous functions is totally bounded in the hypo-distance. Thus, numerous technical challenges including those related to existence of estimators and their cluster points are more appropriately and easily addressed. Convergence in this metric is well understood and visualized as the convergence of the sets of points below the graphs of densities. The choice of metric guarantees *convergence of modes, near-modes, height of modes, and high-likelihood events* not necessarily achieved by convergence in the Hellinger distance, pointwise convergence, and related metrics. Thus, new classes of strongly consistent *plug-in estimators* for such quantities emerge from this article.

With our focus on constraints, it is essential that the chosen metric on families of densities facilitates

the formulation of specific constraints. It would also be useful if subsets of densities specified by constraints were compact. We demonstrate that the hypo-distance indeed is natural and that it achieves the compactness property for the wide variety of constraints and their combinations listed above.

The consideration of this combination of constraints extends the significant body of literature on individual shape constraints going back to [26]; see the recent monograph [28] for a comprehensive treatment and the dissertation [18] for an overview. For example, [30, 32, 43, 19, 4] address univariate log-concave densities with computational comparisons in [52], an extensive review in [60], and results on adaptation in [33]; [17] considers univariate log-concave densities with known mode; [10, 11, 34] deal with multivariate log-concave densities; [29, 41] examine convexity and monotonicity; [42, 41] address monotonicity alone, monotonicity in combination with convexity, U-shape, and unimodality with known mode; [46, 31] are also concerned with unimodal functions, the former dealing with U-shape as well; [8] studies monotonicity, convexity, and log-concavity; [5, 24, 6] examine k-monotonicity; [54] addresses monotone transformations of convex functions; and [44, 45] examine several shapes with a focus on convex problem formulations. For an entry point into the related literature on shape-restricted regression, we refer to [53].

Certain shape constraints combined with sample-size dependent constraints that represent finite-dimensional approximations of a family of densities (sieves) can be found in [15, 14, 13, 45]; see [27, 25, 9] and Theorems 8.4 and 8.12 in [59] for general results on sieve estimators, which typically concentrate on developing approximations of linear spaces using a finite basis. We also consider finite-dimensional approximations, but cannot consider a basis because semicontinuous functions do not form a linear space. Instead, we adopt classes of spline-like approximations (epi-splines) tailored to such spaces. Moreover, we examine sample-size dependent constraints that arise due to incomplete information about shapes and other properties such as confidence intervals for moments as well as due to approximations introduced for computational reasons including inexact evaluation of integrals.

Steps towards addressing more general collections of constraints are taken in [61], where parametric nonlinear least-squares regression is subject to a finite number of smooth equality and inequality constraints. In [50], we examine univariate densities with general constraints in the large parametric class of epi-splines. The closest precursor to the present paper might be [16], which considers consistency using epi-convergence of maximum likelihood estimators of densities on $\mathbb{R}^d$ subject to constraints that form closed sets in some separable Hilbert space. Moreover, either the support of the densities are bounded and the Hilbert space is a reproducing kernel space or all densities are uniformly bounded from above and away from zero. The constraints are fixed and not permitted to vary with the sample size. It is unclear how difficult it is to formulate relevant constraints yielding closed sets in the chosen Hilbert space; few examples are included. Complications also arise from the fact that the set-up provides no guarantee that the estimators will have a cluster point as the sample size tends to infinity. However, if such a cluster point exists then an estimator is shown to be consistent. In the more restrictive case of convex and bounded constraints, cluster points exist in the weak topology and consistency is guaranteed in the sense of that topology. In contrast, we consider families of semicontinuous densities, constraints are allowed to vary with the sample size, and compactness ensures that issues regarding existence of cluster points evaporate even in the nonconvex case.

Regularization and side conditions may, in part, be dealt with through penalties instead of by means of constraints in estimation problems; see the monographs [56, 22] and, for example, [39, 35, 55, 36, 40]. It is often a modeler's choice whether a particular side condition is best represented by a penalty or by a constraint. In this paper, we permit penalties but focus on constraints. Our goal is to enhance the ability to use constraints as modeling tools. We refer to [16] for further discussion about the equivalence and interpretation of reformulation of constraints as penalty terms, which is far from trivial especially in the case of multiple constraints.

In summary, given iid d-dimensional random vectors $X^1, X^2, ..., X^n$, we consider the general class of constrained maximum likelihood density estimators

$$\hat{f}^n \in \varepsilon^n\text{- argmin}_{f \in F^n} -\frac{1}{n}\sum_{j=1}^{n} \log f(X^j) + r^n(f),$$

where $\varepsilon^n \geq 0$ indicates that we may accept near-optimal densities for computational reasons and/or to avoid overfitting, r^n is a penalty function perhaps introduced for the purpose of smoothing (regularizing) the estimates and/or encouraging a unique estimator, and F^n is a class of semicontinuous functions on $\mathbb{R}^d$ that might be tied to the sample size, level of information about shapes and other properties, and various forms of approximations. In fact, F^n might *not only* contain densities, but also their approximations, which could be important when evaluating $\int f(x)dx$ is costly. In this general settings, we show the existence of density estimators and derive their strong consistency as the sample size tends to infinity, knowledge about shapes and other properties improves, and approximations vanish. When constraints are misspecified, i.e., the actual density falls outside these restrictions, we obtain almost sure convergence to a density that is closest to the actual density in Kullback-Leibler divergence.

In addition to examining a wide array of constraints and their combinations, the paper provides a general path to establishing existence and strong consistency by reducing the task to checking that sets formed by constraints are closed with respect to the hypo-distance. This overall approach and concomitant results extend to other criterion functions in density estimation, regression, and classification after minor modifications.

The estimators require numerical solution of optimization problems, which is a nontrivial but well-developed subject. The fact that we permit near-optimal solutions is important because most optimization algorithms only guarantee such solutions; Subsection 4.10 and [50] expand on implementation details. We omit an empirical study as it is already well-known that constraints often have a dramatically positive effect on estimates; see for example [50]. Here, we concentrate on theoretical justifications. Within our framework, extensions to consistency under dependent samples are rather straightforward, but involve technicalities better avoided here and we simply include a remark about this possibility.

Section 2 lays out the framework, especially foundations for semicontinuous functions. Section 3 gives a law of large numbers and the main consistency results. Section 4 considers a series of examples.

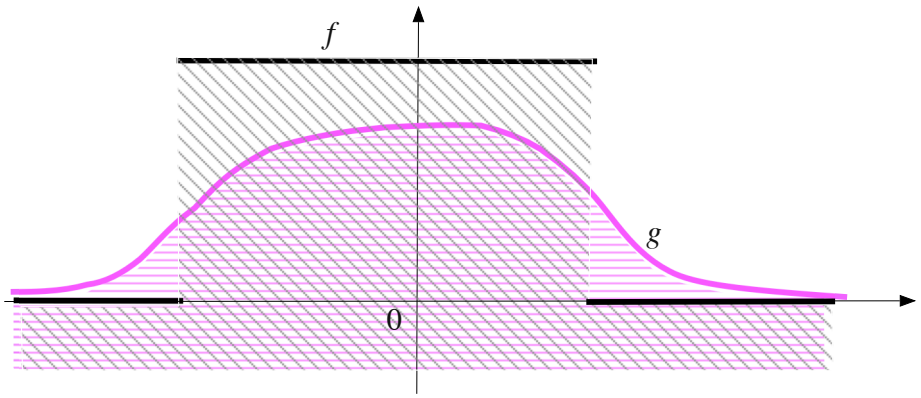

Figure 1: Densities f and g on $S = \mathbb{R}$, with hypographs illustrated by shading.

2 Framework

We start with notation and terminology and denote by $\mathbb{N}$ the positive integers and $\overline{\mathbb{R}} := [-\infty, \infty]$. Given a metric space (Y, d_Y), we recall that a sequence of functions $\{f^n : Y \to \overline{\mathbb{R}},\ n \in \mathbb{N}\}$ *hypo-converge* to a function $f : Y \to \overline{\mathbb{R}}$, denoted by $f^n \xrightarrow{h} f$, when

$$\forall y^n \to y,\ \ \operatorname{limsup} f^n(y^n) \le f(y)$$
$$\forall y, \exists y^n \to y \text{ such that } \operatorname{liminf} f^n(y^n) \ge f(y).$$

Hypo-convergence can be viewed as one-sided uniform convergence as seen later. The sequence *epi-converges* to f, denoted by $f^n \xrightarrow{e} f$, if $-f^n \xrightarrow{h} -f$.

We recall that a function $f : Y \to \overline{\mathbb{R}}$ is *upper semicontinuous* (usc) if for every $y^n \to y$, $\operatorname{limsup} f(y^n) \le f(y)$. It is *lower semicontinuous* (lsc) if $-f$ is usc. In the following, (Y, d_Y) is either $\mathbb{R}^d$ (or a subset thereof) with a metric defined by a norm, or a space of usc functions with the hypo-distance as metric.

2.1 The Space of Upper Semicontinuous Functions

We consider densities that are nonnegative functions on some closed and nonempty set $S \subset \mathbb{R}^d$, possibly equalling the whole $\mathbb{R}^d$. For technical reasons, we assume that $0 \in S$ throughout. Since we permit densities to have value zero on parts of S, the choice of S can differ from the support of the densities of interest and $0 \in S$ is not a significant restriction. For any norm $\|\cdot\|$ on S, we consider the metric space $(S, \|\cdot - \cdot\|)$ and the function space

$$\text{usc-fcns}(S; [0, \infty]) := \{f : S \to [0, \infty] \ :\ f \text{ usc}\},$$

equip with the hypo-distance, which quantifies the distance between usc functions in terms of a distance between their hypographs. Specifically, the *hypograph* of f is $\operatorname{hypo} f := \{(x, x_0) \in S \times \mathbb{R} \ :\ f(x) \ge x_0\}$; see Figure 1. For $S \times \mathbb{R}$, we adopt the product norm $\|(x, x_0)\|_{\mathbb{S}} := \max\{\|x\|, |x_0|\}$, $(x, x_0) \in S \times \mathbb{R}$.

After letting $\operatorname{dist}(\bar{x}, A)$ be the usual point-to-set distance in $S \times I\!R$ under $\|\cdot\|_{\mathbb{S}}$, we are in a position to define, for any $f, g \in \text{usc-fcns}(S; [0, \infty])$, the *hypo-distance* as

$$d\!l(f, g) := \int_0^\infty d\!l_\rho(f, g) e^{-\rho} d\rho,$$

where the *ρ-hypo-distance*

$$d\!l_\rho(f, g) := \max\Big\{ \big| \operatorname{dist}\big(\bar{x}, \operatorname{hypo} f\big) - \operatorname{dist}\big(\bar{x}, \operatorname{hypo} g\big) \big| \ : \ \|\bar{x}\|_{\mathbb{S}} \le \rho \Big\} \ \text{ for } \rho \ge 0.$$

In the case of Figure 1, we see that for $\bar{x} \in S \times I\!R$ near the origin the distance to either hypograph is zero. However, for sufficiently large ρ one needs to consider $\bar{x}$ high above the graphs of f and g where $\operatorname{dist}(\bar{x}, \operatorname{hypo} f)$ is less than $\operatorname{dist}(\bar{x}, \operatorname{hypo} g)$ and $d\!l(f, g) > 0$. Indeed, $(\text{usc-fcns}(S; [0, \infty]), d\!l)$ is a metric space; see [47, Section 7.I] and [48]. We observe that $d\!l_\rho$ and $d\!l$ differ from the Pompeiu-Hausdorff distance, which is essentially only useful in measuring the distance between sets included in a bounded set.

Restrictions to subsets of $\text{usc-fcns}(S; [0, \infty])$ are needed to account for constraints on shapes, integral values, and other properties. Let

$$F \text{ be a nonempty closed subset of } (\text{usc-fcns}(S; [0, \infty]), d\!l)$$

capturing the needs of a particular application; examples are given in Section 4. For the development in Sections 2-3, F is assumed fixed and the analysis takes place on the metric space $(F, d\!l)$, where we also denote by $d\!l$ the restriction of the hypo-distance to F. It is clear that $(F, d\!l)$ is not a linear space, but it possesses other useful properties.

2.1 Proposition (properties of spaces of usc functions) *The metric space $(F, d\!l)$ is compact with $0 \le d\!l(f, g) \le 1$ for $f, g \in F$.*

Proof. We deduce from [47, Theorem 7.58] and [51, Corollary 3.6] that $(F, d\!l)$ is a complete separable metric (Polish) space. Every closed ball in this space is compact as can be deduced from [47, Theorem 7.58]. In view of [48, Proposition 3.1], $0 \le d\!l(f, g) \le 1$ for all $f, g \in \text{usc-fcns}(S; [0, \infty])$. Thus, we also have that the space is totally bounded. □

We note that the resulting *hypo-topology* on F generated by $d\!l$ is indifferent to the choice of the norm $\|\cdot\|$ on S and the following consistency results are not influenced by this choice.

Convergence of sets is central to our development and often provides a geometric interpretation. We recall that the *outer limit* of a sequence of sets $\{A^n\}_{n \in I\!N}$, denoted by $\operatorname{OutLim} A^n$, is the collection of points to which a subsequence of $\{a^n \in A^n, n \in I\!N\}$ converges. The *inner limit*, denoted by $\operatorname{InnLim} A^n$, is the collection of points to which a sequence $\{a^n \in A^n, n \in I\!N\}$ converges. If both limits exist and are identical to A, we say that $\{A^n\}_{n \in I\!N}$ *set-converges* to A and write $A^n \to A$. We retain this terminology for subsets of any metric space[1].

[1]The outer limit is denoted by limsup and the inner limit by liminf in [47] and other references.

In view of Theorem 7.58 in [47] and the preceding discussion, we have the following equivalences for $\{f, f^n, n \in \mathbb{N}\} \subset F$:

$$d\!\!\!l(f^n, f) \to 0 \iff f^n \xrightarrow{h} f \iff \operatorname{hypo} f^n \to \operatorname{hypo} f.$$

We note that the first statement is about convergence of points in the metric space $(F, d\!\!\!l)$, the second one is about hypo-convergence of functions defined on the metric space $(S, \|\cdot - \cdot\|)$, and the third one is about convergence of sets in $(S \times \mathbb{R}, \|\cdot - \cdot\|_{\mathbb{S}})$. Thus, the hypo-distance metrizes hypo-convergence of functions in F and such convergence is easily understood and visualized by set-convergence of hypographs; in Figure 1 the striped region under the graph of g needs to get close to the region under the graph of f for g to have a small hypo-distance from f. We elaborate on connections with uniform, pointwise, and other notions of convergence later, but the following fact is immediate from the definition of hypo-convergence.

2.2 Proposition (relation to pointwise convergence) *If $\{f, f^n, n \in \mathbb{N}\} \subset F$ and $f^n \xrightarrow{h} f$, then:*

(i) For every $x \in S$, $\operatorname{limsup} f^n(x) \leq f(x)$ and actually $f^n(x) \to f(x)$ when $f(x) = 0$.

(ii) For every $x \in S$, there exists a sequence $x^n \in S \to x$ with $f^n(x^n) \to f(x)$.

Since $f \in$ usc-fcns$(S; [0, \infty])$ has closed superlevel sets, f is Borel measurable and its Lebesgue integral is defined, but possibly taking the value infinity. Obvious restrictions are needed. However, a useful feature of our approach is the fact that F might include functions that do *not* integrate to 1. This permits us to consider estimators that only enforce the integral requirement approximately, which is practically important when addressing all but low-dimensional problems.

A premature restriction to densities may also cause technical complications as $f^n \xrightarrow{h} f$ does *not* necessarily imply $\int f^n(x)dx \to \int f(x)dx$ even if all integrals are finite. Indeed, if densities f^n are unchanged for all n, except that they move away from the origin to the horizon and therefore hypo-converge to $f = 0$, then we certainly have $1 = \int f^n(x)dx \not\to \int f(x)dx = 0$. Convergence of integrals may also fail when S is compact[2]. Thus, a restriction of usc-fcns$(S; [0, \infty])$ to densities might construct a space that is not complete. For the same reason, there is no immediate connection between hypo-convergence and integral-type convergence such as those in terms of L_p-norms, Hellinger distance, and Kullback-Leibler divergence. Still, when we restrict the consideration to function with certain shapes and other properties, we usually achieve convergence of integrals automatically; see Section 4.

The hypo-distance is defined relative to the origin in $\mathbb{R}^d \times \mathbb{R}$, which is the reason for insisting on $0 \in S$. Other choices of "center" result in the same hypo-topology on F, but the numerical values of distances between functions may change. In fact, the hypo-distance is not translation invariant. Thus, it may be beneficial to consider some "standardization," for example by translating any data to have zero mean and (marginal) standard deviation one. In the present context, this issue is insignificant as we concentrate on consistency and only the topology matters.

[2]For a counter example, let $\{y^n\}_{n \in N}$ be an enumeration of the rational numbers on $S = [0, 1]$. Let $f^n(x) = 1$ if $x \in \{y^1, y^2, ..., y^n\}$ and $f^n(x) = 0$ otherwise, which makes f^n usc with $\int f^n(x)dx = 0$. However, $f^n \xrightarrow{h} f = 1$ because for $x^n \to x$, $f^n(x^n) \leq f(x)$, which confirms the first condition in the definition of hypo-convergence. For the second condition, let $x \in S$ and $x^n \in \operatorname{argmin}\{|x - y| : y = y^i, i = 1, ...n\}$. Thus, $x^n \to x$ and $f^n(x^n) = 1 = f(x)$ and the claim holds.

For $\delta \geq 0$, we define the maximum value and (near-)maximum points of $f \in F$ by[3]

$$\sup f := \sup\{f(x) \ : \ x \in S\} \ \text{ and } \ \delta\text{- argmax}\, f := \{x \in S \ : \ f(x) \geq \sup f - \delta\}.$$

When $\delta = 0$, we may omit "δ-". If f is a density, then argmax f is the set of *modes* of f and δ- argmax f is the set of *near-modes*. For $\alpha \in [0, \infty)$, the set of *high-likelihood events* is given by the level set

$$\text{lev}_\alpha f := \{x \in S \ : \ f(x) \geq \alpha\}.$$

A consequence of hypo-convergence is the convergence of maximum points and thus also convergence of modes as made specific next. When estimation of modes is important, the present set-up is therefore especially attractive and in fact offers a rich class of estimators for modes. We stress that modes are defined here as *global* maximizers of densities, which deviates from the common terminology that also includes *local* maximizers. Extension to this more inclusive definition of modes is possible but requires additional concepts avoided in this paper.

A strengthened notion of hypo-convergence rules out pathological behavior of modes. We say that a sequence of functions $\{f^n : S \to [0, \infty], n \in I\!N\}$ *hypo-converges tightly* to a function $f : S \to [0, \infty]$ when $f^n \xrightarrow{h} f$ and for all $\varepsilon > 0$, one can find a compact set $B_\varepsilon \subset S$ and an index n_ε such that

$$\forall\, n \geq n_\varepsilon : \quad \sup\{f^n(x) \ : \ x \in S \cap B_\varepsilon\} \geq \sup f^n - \varepsilon.$$

If S is compact, then tightness follows but the condition holds more generally as it essentially only requires that the modes for the whole collection $\{f^n\}_{n \in N}$ is contained in a bounded subset of S.

2.3 Proposition (modes and high-likelihood events) *For $\{f, f^n, n \in I\!N\} \subset F$, $f^n \xrightarrow{h} f$ implies:*

(i) $\liminf(\sup f^n) \geq \sup f$.

(ii) If $\{\delta, \delta^n \geq 0, n \in I\!N\}$ and $\delta^n \to \delta$, then $\text{OutLim}(\delta^n\text{- argmax}\, f^n) \subset \delta\text{- argmax}\, f$.

(iii) For all $\alpha \in [0, \infty]$, $\text{OutLim}(\text{lev}_\alpha f^n) \subset \text{lev}_\alpha f$.

(iv) If $\bar{x}$ is a cluster point of a sequence $\{x^n \in \text{argmax}\, f^n, n \in I\!N\}$, i.e., the limit of a subsequence $\{x^{n_k}\}_{k \in N}$, then $\sup f^{n_k} \to \sup f$ *as* $k \to \infty$.

(v) If in addition $\sup f < \infty$, *then*

$$\begin{aligned} &f^n \xrightarrow{h} f \text{ tightly} \iff \sup f^n \to \sup f \\ &\implies \exists\{\delta^n \downarrow 0, n \in I\!N\} \text{ such that } \delta^n\text{-argmax}\, f^n \to \text{argmax}\, f. \end{aligned}$$

Proof. The argument is given in the appendix. □

[3]Throughout we use the common extended real-valued calculus: $0 \cdot \infty = 0$, $\alpha \cdot \infty = \infty$ for $\alpha > 0$, $\alpha + \infty = \infty$ for $\alpha \in I\!R$, etc. In addition, we adopt the convention that $-\infty + \infty = \infty$.

In view of the definition of the outer limit, part (ii) implies that when all the functions are densities, any cluster point of a sequence of near-modes of f^n is a near-mode of f. If $\delta = 0$, then they are actually modes. Parts (iii)-(v) ensure similar convergence of high-likelihood events and the height of the modes. Convergence of densities in the sense of L_1, L_2, Hellinger, and Kullback-Leibler as well as pointwise convergence fails to ensure convergence of modes, near-modes, high-likelihood events, and height of modes without additional assumptions.

2.2 Estimators

Next, we define the broad class of constrained maximum likelihood estimators under consideration and establish their existence. In addition to independent random vectors $X^1, X^2, ...$ identically distributed as the *actual density* f^0, we consider information about constraints and approximations, which might evolve, possibly in coordination with the sample size in a manner similar to the construction of sieves; see for example [27, 25, 9]. Thus, we also consider as given a sequence of closed subsets $F^1, F^2, ...$ of $(F, d\!\!l)$ that represent this information. We can view these sets as random and revealed with the sample. This involves no complication except additional terminology. Thus, to keep the presentation simple, we view these sets as deterministic. When considering the product space $S \times F$, we adopt the product metric and the product sigma-algebra formed by the Borel sets on both. We extend the logarithm function to $[0, \infty]$ by setting $\log 0 = -\infty$ and $\log \infty = \infty$. For any $\varphi : F \to \overline{I\!R}$, $\varepsilon \geq 0$, and $G \subset F$, we let $\inf \varphi := \inf\{\varphi(f) : f \in F\}$, $\inf_G \varphi := \inf\{\varphi(f) : f \in G\}$,

$$\varepsilon\text{- argmin}\, \varphi := \{f \in F : \varphi(f) \leq \inf \varphi + \varepsilon\}, \text{ and}$$
$$\varepsilon\text{- argmin}_{f \in G}\, \varphi(f) := \{f \in G : \varphi(f) \leq \inf_G \varphi + \varepsilon\}.$$

Given $X^1, X^2, ..., X^n$, a nonempty closed set $F^n \subset F$, an optimality tolerance $\varepsilon^n \geq 0$, and a penalty function $r^n : F \to [0, \infty)$, we define the

$$\text{Estimation Problem:} \quad \text{find } \hat{f}^n \in \varepsilon^n\text{- argmin}_{f \in F^n} -\frac{1}{n}\sum_{j=1}^{n} \log f(X^j) + r^n(f).$$

The remainder of the paper deals with the existence and, most prominently, consistency of estimators $\hat{f}^n$. Technical properties of the (effective) objective function in such problems are needed.

2.4 Lemma (lsc and measurability of integrands) *For closed $G \subset F$ and lsc $r : F \to [0, \infty)$, a function $\psi : S \times F \to \overline{I\!R}$ defined by*

$$\psi(x, f) = -\log f(x) + r(f) \text{ if } f \in G \text{ and } \psi(x, f) = \infty \text{ otherwise,}$$

is lsc and measurable.

Proof. Suppose that $f^n \xrightarrow{h} f$ and $x^n \in S \to x$, then $\operatorname{limsup} f^n(x^n) \leq f(x)$ by the definition of hypo-convergence. Thus, $\operatorname{liminf} -\log f^n(x^n) \geq -\log f(x)$, which implies that the function $(x, f) \mapsto -\log f(x)$ is lsc. Since r is also lsc and G is closed, ψ is lsc. Measurability then follows directly from the fact that sublevel sets are closed. □

2.5 Lemma (additional properties of integrands) *If $f^n \xrightarrow{h} f$ implies $f^n(x) \to f(x), x \in S$, for functions in F, then $\psi : S \times F \to \overline{\mathbb{R}}$ defined by $\psi(x,f) = \log f(x)$ is usc and measurable, and the functions $\{\psi(x,\cdot), x \in S\}$ are lsc.*

Proof. When $f^n \xrightarrow{h} f$ and $x^n \in S \to x$, limsup $f^n(x^n) \le f(x)$ and limsup $\log f^n(x^n) \le \log f(x)$, which ensures that ψ is usc. Its measurability is then immediate from the fact that superlevel sets are closed. For $x \in S$, $f^n(x) \to f(x)$ implies that liminf $\log f^n(x) \ge \log f(x)$. Thus, $\psi(x,\cdot)$ is lsc. □

2.6 Theorem (existence of estimators) *If $F^n \subset F$ is nonempty and closed, $r^n : F \to [0,\infty)$ is lsc, $\varepsilon^n \ge 0$, $\{x^j\}_{j=1}^n \subset S$, and $f(x^j) < \infty$ for all $j = 1,...,n$ and $f \in F^n$, then the solution set*

$$\varepsilon^n\text{-}\operatorname{argmin}_{f \in F^n} -\frac{1}{n}\sum_{j=1}^n \log f(x^j) + r^n(f) \text{ is nonempty and closed.}$$

Proof. The objective function is lsc by Lemma 2.4 and the fact that the sum of lsc functions is lsc provided that the summands are greater than $-\infty$. The conclusion then follows from the compactness of F^n. □

Although the general subject of existence of a maximum likelihood estimator is nontrivial (see for example [7] for the unimodal case), we here achieve existence with ease due to the choice of metric space $(F, d\!l)$, which is always compact. The solution sets of the Estimation Problem are rather involved and may not be singletons as can be expected in this general context with nearly arbitrary constraints. The situation resembles that in high-dimensional statistics where the nature of an estimate is discovered and also not known a priori. As in those settings, the penalty term r^n can be used to ensure uniqueness and favor sparsity of an estimator.

3 Strong Consistency

We now turn to the main consistency results and view the Estimator Problem as an approximation of the Actual Problem:

$$\text{find } f^\star \in \operatorname{argmin}_{f \in F^0} -E[\log f(X^1)] := -\int \log f(x) f^0(x) dx,$$

where $F^0 \subset F$ is a nonempty set that typically contains the actual density f^0, but under misspecification of constraints may not. If F^0 contains only densities and the actual density f^0, then $f^0 \in \operatorname{argmin}_{f \in F^0} -E[\log f(X^1)]$ and all elements $f^\star$ of this argmin are *equivalent* in the sense that they deviate from f^0 at most on a set of Lebesgue measure. Thus, the issue of consistency reduces to that of establishing almost sure set-convergence of the near-optimal solutions $\varepsilon^n\text{-}\operatorname{argmin}_{f \in F^n} -(1/n)\sum_{j=1}^n \log f(X^j) + r^n(f)$ to the optimal solutions $\operatorname{argmin}_{f \in F^0} -E[\log f(X^1)]$.

3.1 Epi-Convergence

Epi-convergence is the principal tool for establishing convergence of optimal solutions of minimization problems; see for example [47, 49, Chapter 7]. Section 2 hints to the key results for hypo-convergence and its consequence for maximization problems. Here, we recall the central facts for minimization problems and develop an intermediate result.

3.1 Proposition (consequences of epi-convergence) *Suppose that the functions $\varphi^n : F \to \overline{I\!R}$ epi-converge to $\varphi : F \to (-\infty, \infty]$ and $\varphi \not\equiv \infty$. Then,*

(i) $\inf \varphi^n \to \inf \varphi \in I\!R$

(ii) $\forall \{\varepsilon^n \downarrow 0, n \in I\!N\}, \mathrm{OutLim}(\varepsilon^n\text{-}\operatorname{argmin} \varphi^n) \subset \operatorname{argmin} \varphi \neq \emptyset$

(iii) $\exists \{\varepsilon^n \downarrow 0, n \in I\!N\}$ *such that* $\varepsilon^n\text{-}\operatorname{argmin} \varphi^n \to \operatorname{argmin} \varphi$.

Proof. These results and arguments are essentially in [47, Chapter 7] and [49], but here specialized due to the compactness of F. It is immediate that $\inf \varphi < \infty$. Since $\varphi^n \xrightarrow{e} \varphi$, φ must be lsc [47, Proposition 7.4(a)]. Thus, the compactness of F ensures that $\operatorname{argmin} \varphi \neq \emptyset$. Since $\varphi > -\infty$, it follows that $\inf \varphi > -\infty$. This establishes the right-most conclusions in parts (i) and (ii).

Next, we consider part (i) and first establish that $\operatorname{limsup}(\inf \varphi^n) \leq \inf \varphi$. There exist $f \in \operatorname{argmin} \varphi$ and $f^n \xrightarrow{h} f$ such that $\operatorname{limsup} \varphi^n(f^n) \leq \varphi(f)$. Thus, $\operatorname{limsup}(\inf \varphi^n) \leq \operatorname{limsup} \varphi^n(f^n) \leq \varphi(f) = \inf \varphi$ and the claim is established. Second, we prove that $\operatorname{liminf}(\inf \varphi^n) \geq \inf \varphi$. Suppose for the sake of a contradiction that $\inf \varphi^{n_k} = -\infty$ for some subsequence $(n_1, n_2, ...)$. Then, there exists f^n such that $\varphi^n(f^n) \to -\infty$. The compactness of F ensures that there exists a cluster point f of $\{f^n\}_{n \in N}$, but then epi-convergence φ^n to φ implies that $-\infty = \operatorname{liminf} \varphi^n(f^n) \geq \varphi(f)$, which contradicts the finiteness of $\varphi(f)$. Let $\varepsilon > 0$. Since we have ruled out that $\inf \varphi^n = -\infty$ for sufficiently large n, there exists $f^n \in \varepsilon\text{-}\operatorname{argmin} \varphi^n$ for such n. Let f be a cluster point of $\{f^n\}_{n \in N}$, which exists due to the compactness of F. Thus, due to $\varphi^n \xrightarrow{e} \varphi$,

$$\operatorname{liminf}(\inf \varphi^n) \geq \operatorname{liminf} \varphi^n(f^n) - \varepsilon \geq \varphi(f) - \varepsilon \geq \inf \varphi - \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $\operatorname{liminf}(\inf \varphi^n) \geq \inf \varphi$ and part (i) is established.

Following the same arguments as those in the proof of part (i) of Proposition 2.3, we obtain part (ii). Part (iii) follows by a direct application of [49, Theorem 3.8]. □

The following proposition is helpful in establishing epi-convergence under approximations of constraint sets, where we use the notation $\iota_G : F \to [0, \infty]$ for the function that assigns zero on $G \subset F$ and infinity elsewhere.

3.2 Proposition (epi-convergence under constraint approximations) *For $\varphi, \varphi^n : F \to \overline{I\!R}$ and $F^0, F^n \subset F$, suppose that $\varphi^n \xrightarrow{e} \varphi$, $\varphi^n \xrightarrow{h} \varphi$, and $F^n \to F^0$. Then, $\varphi^n + \iota_{F^n} \xrightarrow{e} \varphi + \iota_{F^0}$.*

Proof. Let $f^n \xrightarrow{h} f$. The epi- and hypo-convergence of φ^n to φ imply that $\varphi^n(f^n) \to \varphi(f)$. If $f \in F^0$, then $\liminf \varphi^n(f^n) + \iota_{F^n}(f^n) \geq \varphi(f) = \varphi(f) + \iota_{F^0}(f)$. Otherwise, if $f \notin F^0$, then, in view of the fact that $F^n \to F^0$, there exists $\bar{n}$ such that for all $n \geq \bar{n}$, $f^n \notin F^n$. Thus, again $\liminf \varphi^n(f^n) + \iota_{F^n}(f^n) \geq \varphi(f) + \iota_{F^0}(f)$.

Let $f \in F^0$. Since $F^n \to F^0$, there exists $f^n \in F^n \xrightarrow{h} f$. Since $\varphi^n \xrightarrow{h} \varphi$, we have that $\limsup \varphi^n(f^n) + \iota_{F^n}(f^n) \leq \varphi(f) = \varphi(f) + \iota_F(f)$. The same inequality trivially holds for $f \notin F^0$. □

3.2 Law of Large Numbers

We rely on a functional law of large numbers, essentially in [2, 38], that ensures almost sure epi-convergence of sample average functions. For completeness, we include the statement as well as a new proof, which is simpler than that in [2]. It follows the arguments in [38] for ergodic processes, but takes advantage of the present iid setting. The result of this subsection will be used for the metric space $(F, d\!\!l)$ and the probability space generated by the random vector X^1, but no complications arise from making the statements somewhat more general. Suppose that (Y, d_Y) is a complete separable metric (Polish) space, with Borel sigma-algebra $\mathcal{B}_Y$, and $(\Xi, \mathcal{A}, P)$ is a complete probability space. As is apparent from [38], the following theorem holds even without the completeness of $(\Xi, \mathcal{A}, P)$. However, this introduces some technical details better avoided here.

A function $\psi : \Xi \times Y \to \overline{I\!R}$ is said to be *random lsc* if it is $(\mathcal{A} \otimes \mathcal{B}_Y)$-measurable and for all $\xi \in \Xi$, $\psi(\xi, \cdot)$ is lsc. It is also *locally inf-integrable* if[4]

$$\forall y \in Y \ \exists V, \text{ a closed neighborhood of } y, \text{ s.t. } \int \inf_V \psi(\xi, \cdot) dP(\xi) > -\infty.$$

We note that by [47, Theorem 14.37], the function $\xi \mapsto \inf_V \psi(\xi, \cdot)$ is measurable. A benefit now emerges from our reliance on epi-convergence instead of uniform convergence of the criterion functions: The following functional law of large numbers only requires local inf-integrability (i.e., essentially a local lower bound on the integrand), which implies that the integrand can be arbitrarily high and even take the value ∞. In contrast, if the goal were uniform convergence of the criterion functions, then the integrand would have had to be uniformly bounded from below *and* above by an integrable function.

3.3 Proposition (law of large numbers) *Suppose that (Y, d_Y) is a complete separable metric space, $(\Xi, \mathcal{A}, P)$ is a complete probability space, and $\psi : \Xi \times Y \to \overline{I\!R}$ is a locally inf-integrable random lsc function. If $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, ...$ is a sequence of independent random elements that take values in Ξ with distribution P, then*

$$\frac{1}{n} \sum_{j=1}^{n} \psi(\boldsymbol{\xi}^j, \cdot) \xrightarrow{e} E[\psi(\boldsymbol{\xi}^1, \cdot)] \text{ a.s.}$$

Proof. We start by showing that $E[\psi(\boldsymbol{\xi}^1, \cdot)]$ is lsc and let $y^n \to y$. Since ψ is locally inf-integrable and $\psi(\xi, \cdot)$ is lsc, a slight extension of Fatou's Lemma (see [16, Appendix]) ensures that

$$\liminf \int \psi(\xi, y^n) dP(\xi) \geq \int \Big(\liminf \psi(\xi, y^n) \Big) dP(\xi) \geq \int \psi(\xi, y) dP(\xi)$$

[4]For measurable $h : \Xi \to \overline{I\!R}$, $\int h(\xi) dP(\xi) = \int \max\{0, h(\xi)\} dP(\xi) - \int \max\{0, -h(\xi)\} dP(\xi)$, with $\infty - \infty = \infty$.

and the claim is established.

Let $\bar{D} \subset Y \times \overline{I\!R}$ be a countable dense subset of the epigraph epi $E[\psi(\boldsymbol{\xi}^1, \cdot)]$, with epi $h := \{(y, y_0) \in Y \times I\!R \ : \ h(y) \leq y_0\}$. Moreover, let $D \subset Y$ be a countable dense subset of Y that contains the projection of $\bar{D}$ on Y and Q_+ be the nonnegative rational numbers. For $y \in D$ and $r \in Q_+$, we define $\pi_{y,r} : \Xi \to \overline{I\!R}$ by setting

$$\pi_{y,r}(\xi) := \inf_{\boldsymbol{B}^o(y,r)} \psi(\xi, \cdot) \text{ if } r > 0 \text{ and } \pi_{y,0}(\xi) := \psi(\xi, y),$$

where $I\!B^o(y, r) := \{y' \in Y \ : \ d_Y(y', y) < r\}$. By Theorem 3.4 in [38], every such $\pi_{y,r}$ is an extended real-valued random variable defined on the probability space $(\Xi, \mathcal{A}, P)$. Since ψ is locally inf-integrable, it follows that for every $y \in D$ there exists a closed neighborhood V_y of y and $r_y \in (0, \infty)$ such that

$$I\!B^o(y, r) \subset V_y \text{ and } E[\pi_{y,r}] \geq \int \inf_{V_y} \psi(\xi, \cdot) dP(\xi) > -\infty \text{ for } r \in [0, r_y].$$

Let $(\Xi^\infty, \mathcal{A}^\infty, P^\infty)$ be the product space constructed from $(\Xi, \mathcal{A}, P)$ in the usual manner. For every $y \in D$ and $r \in [0, r_y] \cap Q_+$, a standard law of large numbers for extended real-valued random variables (see for example [21, Theorems 7.1 and 7.2]) ensures that

$$\frac{1}{n} \sum_{j=1}^{n} \pi_{y,r}(\xi^j) \to E[\pi_{y,r}] \text{ for } P^\infty\text{-a.e. } (\xi^1, \xi^2, ...) \in \Xi^\infty.$$

Since $\{\pi_{y,r} : y \in D, r \in [0, r_y] \cap Q_+\}$ is a countable collection of random variables, there exists $\Xi_0^\infty \subset \Xi^\infty$ such that $P(\Xi_0^\infty) = 1$ and

$$\frac{1}{n} \sum_{j=1}^{n} \pi_{y,r}(\xi^j) \to E[\pi_{y,r}] \text{ for all } (\xi^1, \xi^2, ...) \in \Xi_0^\infty \text{ and } y \in D, r \in [0, r_y] \cap Q_+.$$

We proceed by addressing the two part of the definition of epi-convergence. First, suppose that $y^n \to y$. There exist $\bar{n}^k \in I\!N$, $z^k \in D$, and $r^k \in [0, r_y] \cap Q_+$, $k \in I\!N$, such that $z^k \to y$, $r^k \to 0$,

$$I\!B^o(z^k, r^k) \supset I\!B^o(z^{k+1}, r^{k+1}), \text{ and } y^n \in I\!B^o(z^k, r^k) \text{ for } n \geq \bar{n}^k, k \in I\!N.$$

We temporarily fix k. Then, for $n \geq \bar{n}^k$ and $(\xi^1, \xi^2, ...) \in \Xi_0^\infty$,

$$\frac{1}{n} \sum_{j=1}^{n} \psi(\xi^j, y^n) \geq \frac{1}{n} \sum_{j=1}^{n} \inf_{\boldsymbol{B}^o(z^k, r^k)} \psi(\xi^j, \cdot) = \frac{1}{n} \sum_{j=1}^{n} \pi_{z^k, r^k}(\xi^j) \to E[\pi_{z^k, r^k}].$$

The nestedness of the balls, implies that $\pi_{z^k, r^k} \leq \pi_{z^{k+1}, r^{k+1}}$ for all k. Moreover the lsc of $\psi(\xi, \cdot)$ implies that for all $\xi \in \Xi$, $\pi_{z^k, r^k}(\xi) \to \pi_{y,0}(\xi) = \psi(\xi, y)$. Thus, in view of the monotone convergence theorem, $E[\pi_{z^k, r^k}] \to E[\psi(\boldsymbol{\xi}^1, y)]$. We have establish that for $(\xi^1, \xi^2, ...) \in \Xi_0^\infty$, liminf $(1/n) \sum_{j=1}^{n} \psi(\xi^j, y^n) \geq E[\psi(\boldsymbol{\xi}^1, y)]$.

Second, for every $y \in Y$, we construct a sequence $y^n \to y$ such that for $(\xi^1, \xi^2, ...) \in \Xi_0^\infty$, limsup$(1/n) \sum_{j=1}^{n} \psi(\xi^j, y^n) \leq E[\psi(\boldsymbol{\xi}^1, y)]$.

Suppose that $y \in D$. Then, the claim holds because for $(\xi^1, \xi^2, ...) \in \Xi_0^\infty$

$$\operatorname{limsup} \frac{1}{n} \sum_{j=1}^{n} \psi(\xi^j, y) = \frac{1}{n} \sum_{j=1}^{n} \pi_{y,0}(\xi^j) \to E[\pi_{y,0}] = E[\psi(\boldsymbol{\xi}^1, y)].$$

Fix $(\xi^1, \xi^2, ...) \in \Xi_0^\infty$ and let $h : Y \to \overline{I\!R}$ be the unique lsc functions that has as epigraph the set $\operatorname{OutLim} \operatorname{epi}[(1/n) \sum_{j=1}^{n} \psi(\xi^j, \cdot)]$. Thus, the prior equality is equivalent to having $h(y) \leq E[\psi(\boldsymbol{\xi}^1, y)]$, which then holds for all $y \in D$. Consequently, $\{(y, \alpha) \in Y \times I\!R \ : \ h(y) \leq \alpha, y \in D\} \subset \operatorname{epi} E[\psi(\boldsymbol{\xi}^1, \cdot)]$. Since h is lsc and $\operatorname{epi} E[\psi(\boldsymbol{\xi}^1, \cdot)]$ is closed from the earlier established fact that $E[\psi(\boldsymbol{\xi}^1, \cdot)]$ is lsc, we have after taking the closure on both sides that $\operatorname{epi} h \subset \operatorname{epi} E[\psi(\boldsymbol{\xi}^1, \cdot)]$ and also $h(y) \leq E[\psi(\boldsymbol{\xi}^1, y)$ for all y. By construction of h, this implies that for all y there exists $y^n \to y$ such that $\operatorname{limsup}(1/n) \sum_{j=1}^{n} \psi(\xi^j, y^n) \leq E[\psi(\boldsymbol{\xi}^1, y)]$ and the conclusion holds. □

3.3 Main Results: Strong Consistency

We are now in a position to develop consistency results and give two versions; with and without approximation of the constraint set.

3.4 Theorem (consistency under fixed constraints) *Suppose that $X^1, X^2, ...$ are independent random vectors with common density $f^0 : S \to [0, \infty]$, for every $f \in F$ there exists a closed neighborhood G of f s.t. $\int [\sup_{g \in G} \log g(x)] f^0(x) dx < \infty$, and there exists $f \in F$ s.t. $\int [\log f(x)] f^0(x) dx > -\infty$. If $\{r^n : F \to [0, \infty), n \in I\!N\}$ hypo-converge to the zero function on F, then*

(i) $\forall \varepsilon^n \to 0$, $\varepsilon^n \geq 0$,

$$\operatorname{OutLim} \left(\varepsilon^n\text{-}\operatorname{argmin}_{f \in F} -\frac{1}{n} \sum_{j=1}^{n} \log f(X^j) + r^n(f) \right) \subset \operatorname*{argmin}_{f \in F} -E[\log f(X^1)] \text{ a.s.}$$

(ii) $\exists \varepsilon^n \to 0$, $\varepsilon^n \geq 0$,

$$\left(\varepsilon^n\text{-}\operatorname{argmin}_{f \in F} -\frac{1}{n} \sum_{j=1}^{n} \log f(X^j) + r^n(f) \right) \to \operatorname*{argmin}_{f \in F} -E[\log f(X^1)] \text{ a.s.}$$

Proof. The function $\psi : S \times F \to \overline{I\!R}$ given by $\psi(x, f) = -\log f(x)$ is random lsc by Lemma 2.4 and also locally inf-integrable in view of the upper integral bound. Proposition 3.3 applies and $(1/n) \sum_{j=1}^{n} \psi(X^j, \cdot) \xrightarrow{e} E[\psi(X^1, \cdot)]$ a.s. Then, for all $f^n \xrightarrow{h} f$,

$$\operatorname{liminf} \frac{1}{n} \sum_{j=1}^{n} \psi(X^j, f^n) + r^n(f^n) \geq \operatorname{liminf} \frac{1}{n} \sum_{j=1}^{n} \psi(X^j, f^n) \geq E[\psi(X^1, f)] \text{ a.s.}$$

Since $r^n \xrightarrow{h} 0$, $r^n(f^n) \to 0$ for all $f^n \xrightarrow{h} f$. Consequently, for sequences $f^n \xrightarrow{h} f$ that furnish $(1/n) \sum_{j=1}^{n} \psi(X^j, f^n) \to E[\psi(X^1, f)]$ a.s. we also have $(1/n) \sum_{j=1}^{n} \psi(X^j, f^n) + r^n(f^n) \to E[\psi(X^1, f)]$ a.s. and we

have established the epi-convergence $(1/n)\sum_{j=1}^{n}\psi(X^j,\cdot)+r^n \xrightarrow{e} E[\psi(X^1,\cdot)]$ a.s. The lower integral bound ensures that $E[\psi(X^1,\cdot)]$ is not identically equal to infinity on F and the upper integral bound guarantees that it is never $-\infty$. Thus, the conditions of Proposition 3.1 are satisfied and the conclusions follow by a direct application of that result. □

The assumptions of the theorem are quite mild. The upper integral bound is trivially satisfied if all functions in F are bounded from above with a common constant, but also other conditions exist as exemplified in Section 4. The lower integral bound is satisfied with $f = f^0$ provided that $f^0 \in F$ and does not have infinite entropy. Again, numerous other possibilities exist. There is no requirement in Theorem 3.4 that every $f \in F$ is a density.

Under the additional assumptions that F contains only densities and $f^0 \in F$, part (i) of Theorem 3.4 implies that regardless of which near-optimal solutions of the Estimation Problem are selected as estimators, they converge to f^0 or an equivalent density, possibly after passing to a subsequence, almost surely. It is the compactness of F that ensures the existence of a convergent subsequence regardless of the choice of estimator from the solution set of the Estimation Problem.

The actual density f^0 may be outside F as is the case when constraints are misspecified. Theorem 3.4 then implies convergence to a density in F nearest to f^0 in the Kullback-Leibler divergence, which is defined as[5]

$$d^{\mathrm{KL}}(g\|f) := \int g(x)\log\frac{g(x)}{f(x)}dx \text{ for densities } f,g : S \to [0,\infty].$$

These observations are summarized in the following corollary.

3.5 Corollary *Suppose that the assumptions of Theorem 3.4 hold, $\varepsilon^n \downarrow 0$, F contains only densities, and*

$$\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F} -\frac{1}{n}\sum_{j=1}^{n}\log f(X^j) + r^n(f).$$

Then, almost surely, $\{\hat{f}^n\}_{n\in N}$ has at least one cluster point and every such point $f^\star$ satisfies

$$f^\star \in \operatorname{argmin}_{f\in F} d^{\mathrm{KL}}(f^0\|f)$$

and, provided that the actual density $f^0 \in F$, also that $f^\star(x) = f^0(x)$ for all $x \in S$ except possibly on set of Lebesgue measure zero.

Proof. The conclusions follow from the discussion prior to the corollary and the fact that the set of minimizers of

$$d^{\mathrm{KL}}(f^0\|f) = E[\log f^0(X^1)] - E[\log f(X^1)]$$

with respect to $f \in F$ contains $\operatorname{argmin}_{f\in F} -E[\log f(X^1)]$ as $E[\log f^0(X^1)]$ is independent of f. □

Part (ii) of Theorem 3.4 implies that if ε^n tends to zero sufficiently slowly, then there are solutions of the Estimation Problem that converge to f^0 provided that F contains only densities and $f^0 \in F$. This

[5] As usual, $\alpha\log(\alpha/\beta)$ is understood as zero for $\alpha = 0$ regardless of $\beta \in [0,\infty]$ and as ∞ for $\alpha \in (0,\infty]$ and $\beta = 0$.

conclusion establishes that even f^0, and not only those that are equivalent to it, can be approached, at least conceptually, by estimators of this form.

We next turn to the version involving approximation of the constraints. Mainly, there are two reasons for studying such approximations: When developing algorithms for computing estimates, there is often a need for approximating the class of functions in F by simpler ones given by a finite number parameters, replacing integrals in constraints by numerical approximations, and substituting infinite numbers of constraints by finite collections. The second reason arises when studying the evolution of auxiliary information about the actual density, for example about its shape, function values, and moments. This information is often subjective, speculative, and related to model choices. It is therefore of interest to consider consistency as such information improves in the sense that the constraints approximate in an increasingly more accurate manner constraints representing properties of an actual density.

3.6 Theorem (consistency under constraint approximation) *Suppose that $X^1, X^2, \ldots$ are independent random vectors with common density $f^0 : S \to [0,\infty]$, the closed nonempty sets $F^0, F^n \subset F$ satisfy $F^n \to F^0$, and for every $f \in F$ there exists a closed neighborhood G of f such that $\int[\sup_{g\in G} \log g(x)]f^0(x)dx < \infty$ and $\int[\inf_{g\in G} \log g(x)]f^0(x)dx > -\infty$. If $\{r^n : F \to [0,\infty), n \in \mathbb{N}\}$ hypo-converge to the zero function and for functions in F, $f^n \xrightarrow{h} f$ implies $f^n(x) \to f(x), x \in S$, then*

(i) $\forall \varepsilon^n \to 0$, $\varepsilon^n \geq 0$,

$$\operatorname{OutLim}\left(\varepsilon^n\text{-}\operatorname{argmin}_{f\in F^n} -\frac{1}{n}\sum_{j=1}^{n}\log f(X^j) + r^n(f)\right) \subset \operatorname*{argmin}_{f\in F^0} -E[\log f(X^1)] \text{ a.s.}$$

(ii) $\exists \varepsilon^n \to 0$, $\varepsilon^n \geq 0$,

$$\left(\varepsilon^n\text{-}\operatorname{argmin}_{f\in F^n} -\frac{1}{n}\sum_{j=1}^{n}\log f(X^j) + r^n(f)\right) \to \operatorname*{argmin}_{f\in F^0} -E[\log f(X^1)] \text{ a.s.}$$

Proof. The function $\psi : S \times F \to \overline{\mathbb{R}}$ given by $\psi(x,f) = -\log f(x)$ is random lsc by Lemma 2.4 and also locally inf-integrable in view of the upper integral bound. Since $f^n \xrightarrow{h} f$ implies pointwise convergence on F, Lemma 2.5 ensures that $-\psi$ is also a locally inf-integrable random lsc because of the lower integral bound. Thus, Proposition 3.3 applies and

$$\frac{1}{n}\sum_{j=1}^{n}\psi(X^j,\cdot) \xrightarrow{e} E[\psi(X^1,\cdot)] \text{ and } \frac{1}{n}\sum_{j=1}^{n}\psi(X^j,\cdot) \xrightarrow{h} E[\psi(X^1,\cdot)] \text{ a.s.}$$

Following a similar argument as in the proof of Theorem 3.4, we find that these results hold after adding r^n to the sample averages. In view of Proposition 3.2 and the fact that $F^n \to F^0$,

$$\frac{1}{n}\sum_{j=1}^{n}\psi(X^j,\cdot) + r^n + \iota_{F^n} \xrightarrow{e} E[\psi(X^1,\cdot)] + \iota_{F^0} \text{ a.s.}$$

where $\iota_G : F \to [0,\infty]$ is zero on $G \subset F$ and infinity otherwise. The lower integral bound ensures that $E[\psi(X^1,\cdot)]$ is not identically equal to infinity on F and the upper integral bound guarantees that it is never $-\infty$. Thus, Proposition 3.1 applies and yields the result. □

The assumptions in the theorem are somewhat more stringent than those of Theorem 3.4; the lower integral bound is strengthened and now requires that densities have values not "too much" near zero as exemplified in Section 4. The requirement that hypo-convergence on F implies pointwise convergence is illustrated in Section 4. Generally, the discussion after Theorem 3.4 carries over to the present case with minor changes. In particular, the following corollary holds in view of the arguments leading to Corollary 3.5.

3.7 Corollary *Suppose that the assumptions of Theorem 3.6 hold, $\varepsilon^n \downarrow 0$, F^0 contains only densities, and*

$$\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F^n} -\frac{1}{n}\sum_{j=1}^{n} \log f(X^j) + r^n(f),$$

Then, almost surely, $\{\hat{f}^n\}_{n\in\mathbb{N}}$ has at least one cluster point and every such point $f^\star$ satisfies

$$f^\star \in \operatorname{argmin}_{f\in F^0} d^{\mathrm{KL}}(f^0\|f)$$

and, provided that the actual density $f^0 \in F^0$, also that $f^\star(x) = f^0(x)$ for all $x \in S$ except possibly on set of Lebesgue measure zero.

As we rely on Proposition 3.3, the above results are limited to iid samples. However, the same results hold provided that, in the notation of the proof of Theorem 3.4, $(1/n)\sum_{j=1}^{n} \psi(X^j,\cdot) \xrightarrow{e} E[\psi(X^1,\cdot)]$ a.s. remains valid. There are many situations where such a strong law of large numbers hold including the one described in [38] for ergodic processes; see also [58, 1]. Thus, we provide a path to more general strong consistency results too.

3.4 Strong Consistency of Plug-In Estimators

Among the many plug-in estimators that can be derived from a density estimator, those of modes, near-modes, height of modes, and high-likelihood events are especially accessible within our framework because strong consistency is *automatically* inherited from that of the density estimator.

3.8 Theorem (plug-in estimators of modes and related quantities) *Suppose that density estimators $\hat{f}^n \xrightarrow{h} f^0$ almost surely. If $\{\delta, \delta^n \geq 0,\ n \in \mathbb{N}\}$, $\delta^n \to \delta$, and $\alpha \in [0,\infty]$, then the plug-in estimators*

$$\hat{m}^n \in \delta^n\text{-}\operatorname{argmax} \hat{f}^n \quad \text{and} \quad \hat{l}^n \in \operatorname{lev}_\alpha \hat{f}^n$$

are strongly consistent in the sense that every cluster point of $\{\hat{m}^n\}_{n\in\mathbb{N}}$ is in δ-$\operatorname{argmax} f^0$, and every cluster point of $\{\hat{l}^n\}_{n\in\mathbb{N}}$ is in $\operatorname{lev}_\alpha f^0$ almost surely.

Moreover, if $\hat{f}^n \xrightarrow{h} f^0$ tightly, almost surely, then the plug-in estimator

$$\hat{h}^n = \sup \hat{f}^n$$

is strongly consistent, i.e., $\hat{h}^n \to \sup f^0$ almost surely.

Proof. Since $\hat{f}^n \xrightarrow{h} f^0$ a.s., the results follow immediately from Proposition 2.3. □

The theorem provides foundations for a rich class of *constrained* estimators for modes, near-modes, height of modes, and high-likelihood events. For example, in Section 4.9 we obtain consistent estimators for these quantities that account for information about moments, support, concavity, and size of subgradients. We observe that the theorem holds even if f^0 fails to have a unique mode.

4 Applications

This section illustrates some possibilities, addressing classical as well as novel constraints. In each example, we follow similar paths. If there is no constraint approximation, Theorem 3.4 (Corollary 3.5) is the main vehicle and we construct a closed set $F \subset \text{usc-fcns}(S; [0, \infty])$ of densities with the desired shapes and properties and for which the upper and lower integral bounds hold.

When addressing constraint approximations, for example because the estimates will be taken from a family of functions *not* containing the actual density, then we invoke Theorem 3.6 (Corollary 3.7). The closed set F contains functions satisfying the stronger upper and lower integral bounds and the pointwise convergence requirement. It may contain non-densities. The closed set $F^0 \subset F$ contains densities with the shapes and properties under consideration. Finally, we construct the closed sets $F^n \subset F$ as approximations of F^0. They may contain non-densities, but tend to F^0 so that convergence to f^0 is possible provided that $f^0 \in F^0$.

Although we usually consider just a few shapes and/or conditions at the time, it is easy to analyze more complex cases using these examples as building blocks and the fact that the intersection of closed sets is closed. This simple fact narrows the task of obtaining consistency of a large collection of estimators to confirming that the set specified by the constraints is closed.

4.1 Hypo-Convergence under Shape Restrictions

We start with general results about shapes of hypo-converging functions being carried over to their limit and also provide connections to pointwise and uniform convergence; proofs are given in the appendix. The propositions in this subsection furnish strong arguments for why formulation of density estimators in spaces of usc functions is indeed innate. We denote by $\operatorname{int} S$ the interior of S.

4.1 Proposition (convexity and (log-)concavity) *For $\{f, f^n, n \in \mathbb{N}\} \subset F$ and $f^n \xrightarrow{h} f$, we have:*

(i) If $\{f^n\}_{n\in\mathbb{N}}$ are concave, then f is concave and for all compact sets $C \subset \operatorname{int} S$, $\sup_{x\in C} |f^n(x) - f(x)| \to 0$. Moreover, if all the functions are finite-valued and for some $\kappa \geq 0$, $\|v\|_2 \leq \kappa$ for every subgradient $v \in \partial f^n(x)$, $x \in S$, then $\|v\|_2 \leq \kappa$ for every $v \in \partial f(x)$, $x \in S$.

(ii) If $\{f^n\}_{n\in\mathbb{N}}$ are log-concave, then f is log-concave and for all compact sets $C \subset \operatorname{int}\{x \in S : f(x) > 0\}$ and $\varepsilon > 0$, there exists $\bar{n}$ such that $e^{-\varepsilon} f(x) \leq f^n(x) \leq e^{\varepsilon} f(x)$ for all $x \in C$ and $n \geq \bar{n}$.

(iii) If $\{f^n\}_{n\in\mathbb{N}}$ are convex and $\operatorname{int} S$ is nonempty, then f is convex.

A log-concave density f is of the form e^g for some concave function g. More general transformations of convex and concave functions lead to the rich class of s-concave densities and beyond; see for example [54, 37].

4.2 Proposition (monotone transformations) *For $\{f, f^n, n \in I\!N\} \subset F$, with $f^n \xrightarrow{h} f$, and a strictly increasing continuous function $h : \overline{I\!R} \to [0, \infty]$, we have:*

(i) If $\{f^n = h(g^n(\cdot)),\ n \in I\!N\}$ with $g^n : S \to \overline{I\!R}$ concave, then $f = h(g(\cdot))$ for some concave function $g : S \to \overline{I\!R}$.

(ii) If $\{f^n = h(g^n(\cdot)),\ n \in I\!N\}$ with $g^n : S \to \overline{I\!R}$ convex and $\operatorname{int} S$ is nonempty, then $f = h(g(\cdot))$ for some convex function $g : S \to \overline{I\!R}$.

Obviously, the transformation by means of strictly decreasing functions, instead of increasing ones, is addressed implicitly by Proposition 4.2.

4.3 Proposition (monotonicity) *For $\{f, f^n, n \in I\!N\} \subset F$ and $f^n \xrightarrow{h} f$, we have:*

(i) If f^n is nondecreasing in the sense that $f^n(x) \le f^n(y)$ for $x \in S$, $y \in \operatorname{int} S$, with $x \le y$ (understood componentwise), then f is also nondecreasing in the same sense.

If S is a box, i.e., $S = [\alpha_1, \beta_2] \times ... [\alpha_d, \beta_d]$, with $-\infty \le \alpha_i < \beta_i \le \infty$, where in the case of $\alpha_i = -\infty$ and $\beta_i = \infty$ the closed intervals are replaced by (half)open intervals, then $\operatorname{int} S$ can be replaced by S.

(ii) If f^n is nonincreasing in the sense that $f^n(x) \ge f^n(y)$ for $x \in \operatorname{int} S$, $y \in S$, with $x \le y$, then f is also nonincreasing in the same sense.

If S is a box, then $\operatorname{int} S$ can be replaced by S.

The limit of a hypo-converging sequence of nondecreasing functions is not necessarily nondecreasing for arbitrary S. Consider $S = \{(x_1, x_2) \in I\!R^2 \ :\ x_1 = x_2, 0 \le x_1, x_2 \le 1\} \cup \{(2, 0)\}$, $f(x) = f^n(x) = 0$ if $x = (2, 0)$, and $f(x) = 1$ and $f^n(x) = \min\{1, n(x_1 + x_2)\}$ otherwise. Clearly, $x = (0, 0) \le y = (2, 0)$, but $f(x) = 1 > f(y) = 0$. Meanwhile, $f^n(x) = f^n(y) = 0$ for all n at these two points and it is nondecreasing elsewhere too. Still, $f^n \xrightarrow{h} f$, as is immediate from the definition.

The nondecreasing property is not sufficient to ensure pointwise convergence from hypo-convergence even if $S = [0, 1]$ as shown by the counterexample $f(x) = 0$ if $x \in [0, 1/2)$ and $f(x) = 1$ otherwise, and $f^n(x) = 0$ if $x \in [0, 1/2 - 1/n)$, $f^n(x) = (n/2)x - n/4 + 1/2$ if $x \in [1/2 - 1/n, 1/2 + 1/n]$ and $f^n(x) = 1$ otherwise. Pointwise convergence fails at $x = 1/2$.

We recall that $f : S \to [0, \infty]$ is Lipschitz continuous with modulus κ when $|f(x) - f(y)| \le \kappa \|x - y\|$ for all $x, y \in S$.

4.4 Proposition (Lipschitz continuity) *Suppose that $\{f, f^n, n \in I\!N\} \subset F$, $f^n \xrightarrow{h} f$, and $\{f^n\}_{n \in N}$ are Lipschitz continuous with common modulus κ. Then, f is also Lipschitz continuous with modulus κ and for every compact set $C \subset S$, $\sup_{x \in C} |f^n(x) - f(x)| \to 0$.*

4.5 Proposition (pointwise bounds) *For $g : S \to [0,\infty]$ and $h \in$ usc-fcns $(S;[0,\infty])$, if $\{f, f^n, n \in \mathbb{N}\} \subset F$, $f^n \xrightarrow{h} f$, and $g(x) \le f^n(x) \le h(x)$, $n \in \mathbb{N}$, $x \in S$, then $g(x) \le f(x) \le h(x)$, $x \in S$.*

4.2 Monotonicity

The first example is the classic Grenander Estimator [26] for nonincreasing densities on $S = [0,\infty)$, which is already fully understood but is included here as a simple introduction to our framework. In the process, we provide a novel approach to handling the well-known issue of lack of consistency at the origin.

Since we permit unbounded densities, a growth condition is required in our framework, cf. Theorem 3.4. Specifically, for $p \in [0,1)$, $q \in (1,\infty)$, and $\alpha \ge q \exp q$, we set

$$F = \Big\{ f \in \text{usc-fcns}([0,\infty),[0,\infty]) \ : \ \int_0^\infty f(x)dx = 1, \\ f(x) \le \min\{\alpha x^{-p}, \alpha x^{-q}\}, \ f(x) \ge f(y) \text{ for } 0 \le x \le y < \infty \Big\}.$$

Since α can be set arbitrarily high, the upper bound on densities in F represents a minor addition to the classical estimator. However, if desirable, the upper bound can be lowered and provides a means to include additional information and assumptions about the actual density.

For $\{f, f^n, n \in \mathbb{N}\} \subset F$, we have that $f^n \xrightarrow{h} f$ implies that $f^n(x) \to f(x)$ provided that f is continuous at $x \in S$, which is seen by the following argument. The definition of hypo-convergence immediately establishes $\operatorname{limsup} f^n(x) \le f(x)$. For $\varepsilon > 0$, there exists $z \in (x,\infty)$ such that $f(x) \le f(z) + \varepsilon$. Hypo-convergence implies that there exist $\bar{n}$ and $z^n \in S \to z$ such that $f^n(z^n) \ge f(z) - \varepsilon$ and $z^n \ge x$ for all $n \ge \bar{n}$. Thus, by monotonicity, for all such n, $f^n(x) \ge f^n(z^n) \ge f(z) - \varepsilon \ge f(x) - 2\varepsilon$. Since $\varepsilon > 0$ is arbitrary, pointwise convergence at x follows when f is continuous at x.

Since a monotone function has discontinuities at most on a set of Lebesgue measure zero, we find that $f^n \xrightarrow{h} f$ implies pointwise convergence almost everywhere. In view on the upper bound on functions in F, an application of the dominated convergence theorem ensures that $\int f^n(x)dx \to \int f(x)dx$. This fact and Propositions 4.3 and 4.5 give that F is closed and nonempty; the latter is guaranteed by the sufficiently large α. The analysis then proceeds on the metric space $(F, d\!l)$.

Suppose that the actual density $f^0 \in F$ has finite mean. We find that

$$\int_0^1 [\sup_{g\in F} \log g(x)] f^0(x)dx \le \int_0^1 \frac{\alpha}{x^p} \log \frac{\alpha}{x^p} dx < \infty \text{ and}$$
$$\int_1^\infty [\sup_{g\in F} \log g(x)] f^0(x)dx \le \int_1^\infty \frac{\alpha}{x^q} \log \frac{\alpha}{x^q} dx < \infty.$$

Moreover, the exponential density $g(x) = q\exp(-qx)$ is in F; the upper bound does not interfere as is easily verified. Thus,

$$\int_0^\infty [\log q e^{-qx}] f^0(x)dx = -q\int_0^\infty x f^0(x)dx + \log q \int_0^\infty f^0(x)dx > -\infty.$$

and Theorem 3.4 applies. The estimator $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F} -(1/n)\sum_{j=1}^n \log f(X^j)$ exists by Theorem 2.6 and, provided that $\varepsilon^n \searrow 0$, converges a.s. by Corollary 3.5 to f^0 or an equivalent density, possibly after passing to a subsequence. In fact, the equivalent density can deviate from f^0 at most at $x = 0$. This is realized by considering a contradiction. If $x > 0$ and $f^0(x) - f(x) = \varepsilon > 0$ for some equivalent density f, then by the usc of f there is a $\delta > 0$ such that $f(z) \le f(x) + \varepsilon/2$ for all z with $|z - x| \le \delta$. Moreover, for such $z \le x$ the nonincreasing property of f^0 implies that $f^0(z) \ge f^0(x) = f(x) + \varepsilon$, which is a contradiction because it shows that f and f^0 differ on an interval with positive Lebesgue measure. Since one can reverse the roles of f^0 and f, the claim holds.

Further insight is available also. The nonincreasing property implies that $\sup f = f(0)$ for all $f \in F$ and therefore any sequence $f^n \xrightarrow{h} f$ in F also converges tightly. Parts (i) and (iii) of Proposition 2.3 then imply that $f^n(0) \to f(0)$. Consequently, possibly after passing to a subsequence, $\hat{f}^n(0) \to f(0)$ a.s. for any density f equivalent to f^0. Thus, we have pointwise convergence a.s. at the origin, but possibly to an equivalent density and not f^0. The usc of such f implies that estimates may have an upward bias. Of course, this is the same predicament as discussed in [62, 3]; the classical Grenander estimator overestimates $f^0(0)$. The former reference provides an elegant remedy involving a penalty term. Here we offer another option that illustrates the easy with which additional constraints can be added to an estimator.

For some $\kappa \in [0, \infty)$, we might consider the redefined

$$F = \Big\{ f \in \text{usc-fcns}([0,\infty),[0,\infty]) \ : \ \int_0^\infty f(x)dx = 1,\ f(0) - f(x) \le \kappa x \\ f(x) \le \min\{\alpha x^{-p}, \alpha x^{-q}\},\ f(x) \ge f(y) \text{ for } 0 \le x \le y < \infty \Big\},$$

where the only change is the introduction of the inequality involving κ. Now all $f \in F$ are continuous at $x = 0$ and also finite at that point. Neither restriction is significant as a discontinuity at zero is of little interest and pointwise convergence at zero is automatic when the actual density is unbounded. Thus, suppose that $f^0 \in F$. The set F is still closed because for $f^n \xrightarrow{h} f$ and $\varepsilon > 0$, there exist $x^n \to 0$ and $\bar{n}$ such that $f^n(x^n) \ge f(0) - \varepsilon/2$ and $f^n(x) \le f(x) + \varepsilon/2$ for all $n \ge \bar{n}$. Consequently, using the nonincreasing property of f^n,

$$f(0) - f(x) \le f^n(x^n) - f^n(x) + \varepsilon \le f^n(0) - f^n(x) \le \kappa x,$$

which establishes the claim that F is closed. Again Theorem 3.4 can be brought in leading to the same conclusions as earlier. Now, the Actual Problem has a unique solution and $\hat{f}^n \xrightarrow{h} f^0$, due to the compactness of F, and $\hat{f}^n(0) \to f^0(0)$ a.s. These conclusions hold for any $\varepsilon^n \searrow 0$.

4.3 Log-concavity

To further illustrate the framework in a well-known setting, we consider the log-concave densities on $S = \mathbb{R}^d$ (see for example [60, 12, 10]). A main purpose of the subsection is to establish the fact that restriction to log-concave densities determines a closed subset of $(\text{usc-fcns}(S;[0,\infty]), d\!\!\!/)$. Thus, we provide the

foundation for establishing consistency of estimators involving this restriction *in combination with others.* We give several examples below of such additional constraints that also determine closed subsets.

For some $\alpha > 0$ and $\beta \in \mathbb{R}$, set

$$F = \Big\{ f \in \text{usc-fcns}(\mathbb{R}^d, [0,\infty]) \ : \ \int f(x)dx = 1, f \text{ log-concave}, \\ f(x) \le \exp(-\alpha\|x\|_1 + \beta) \text{ for } x \in \mathbb{R}^d \Big\}. \tag{1}$$

The upper bound on f complements those derived in a classical setting and is brought in for the purpose of satisfying the assumptions of Theorem 3.4. In view of [12, Lemma 1], all log-concave densities have such an upper bound. The presence of a uniform bound across the whole class eliminates the pathological situations where, for instance, a sequence of Gaussian densities with fixed mean and variances tending to infinity hypo-converges to the zero function as well as a sequence of Gaussian densities with fixed mean and variances tending to zero hypo-converges to a delta-function. Since β can be selected arbitrarily high, the upper bound is not a significant restriction from a practical point of view. However, when beneficial, it can be brought down and be used as a modeling tool to ensure that the estimated density is not too high.

In view of part (i) of Proposition 2.2, $f^n \xrightarrow{h} f$ implies pointwise convergence at x with $f(x) = 0$. Proposition 4.1 ensures pointwise convergence for $x \in \text{int}\{x \in \mathbb{R}^d \ : \ f(x) > 0\}$. Hence, $f^n(x) \to f(x)$ for all $x \in S$ except possibly for $x \notin \text{int}\{x \in \mathbb{R}^d \ : \ f(x) > 0\}$ with $f(x) > 0$. The set of such exceptions has Lebesgue measure zero and the dominated convergence theorem ensures that $\int f^n(x)dx \to \int f(x)dx$ provided that $f^n \xrightarrow{h} f$. This fact and Proposition 4.5 establish that F is closed. Since the intersection with any other closed set (representing other constraints) is also closed, this leads to vast possibilities for building log-concave density estimators with additional restrictions.

Suppose that the actual density $f^0 \in F$. The upper integral bound in Theorem 3.4 is guaranteed through the upper bound on all densities in F. Since log-concave densities have finite entropy due to their finite covariances, the lower integral bound also holds with $f = f^0$. Therefore, Theorem 3.4 applies and the estimator $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F} -(1/n)\sum_{j=1}^n \log f(X^j)$ converges a.s. to the actual density f^0. In this case, the possibility that the Actual Problem has multiple optimal solutions is ruled out due to the fact that two usc, log-concave densities that are identical on $\mathbb{R}^d$, possibly except on a set of Lebesgue measure zero, actually must be identical.

In view of the upper bound on $f \in F$, every convergent sequence in F hypo-converges tightly, which leads to consistency of plug-in estimators of modes, near-modes, height of modes, and high-likelihood events by Theorem 3.8. These results hold even if f^0 fails to have a unique mode. We note that it suffices to consider near-modes $\delta^n\text{-}\operatorname{argmax} \hat{f}^n$ as mode estimators provided that $\delta^n \to 0$.

In [12], we find strong consistency of the maximum likelihood estimator in exponentially weighted total variation norms and, when f^0 is continuous, in exponentially weighted supremum norms. In contrast, we here establish strong consistency in the hypo-distance, i.e., $\hat{f}^n \xrightarrow{h} f^0$ a.s. As discussed above, this implies pointwise convergence possibly except for $x \notin \text{int}\{x \in \mathbb{R}^d \ : \ f^0(x) > 0\}$ with

$f^0(x) > 0$. The convergence is uniform on compact sets $C \subset \text{int}\{x \in \mathbb{R}^d \ : \ f^0(x) > 0\}$ in the sense of Proposition 4.1. Our results hold for any $\varepsilon^n \searrow 0$.

4.4 Lipschitz Continuity

The broad class of Lipschitz continuous densities on $S \subset \mathbb{R}^d$ offers an ability to limit oscillations and overfitting, and has not previously been studied in the context of constrained maximum likelihood density estimation. We let $h : S \to (0, \infty)$ be a bounded usc function, with $\int h(x)dx < \infty$ and $h(x) \geq \exp(-\alpha\|x\|_2^2 + \beta)$, $\alpha > 0$, $\beta \in \mathbb{R}$, and $\kappa \in [0, \infty)$, and define

$$F = \Big\{ f \in \text{usc-fcns}(S, [0, \infty]) \ : \ \int f(x)dx = 1, \ f(x) \leq h(x), \\ |f(x) - f(y)| \leq \kappa\|x - y\|_2, \forall x, y \in S \Big\},$$

which is closed by Prop. 4.4 and 4.5, and the dominated convergence theorem. Again, as in the previous subsection, the upper bound given by h may very well be set high when little information is available about the height of the actual density.

Suppose the actual density $f^0 \in F$ has finite second marginal moments. We invoke Theorem 3.4 for $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F} \ -(1/n)\sum_{j=1}^n \log f(X^j)$. The upper integral bound in the theorem is immediately satisfied since h is bounded from above. For the lower integral bound, we see that there exists a Gaussian $f \in F$, with independent marginals having mean μ and standard deviation σ, and $\int \log f(x) f^0(x)dx = \gamma - \frac{1}{2}\int \sum_{i=1}^d (x_i - \mu)^2/\sigma^2 f^0(x)dx > -\infty$ for some $\gamma \in \mathbb{R}$. Thus, all assumptions of Theorem 3.4 hold.

In this case, the possibility that the Actual Problem has multiple optimal solutions is ruled out due to the fact that two continuous functions that are identical on $\mathbb{R}^d$, possibly except on a set of Lebesgue measure zero, actually must be identical. Consequently, $\hat{f}^n \xrightarrow{h} f^0$ almost surely.

4.5 Location of Modes

We next consider constraints pertaining to information about the location of modes of the actual density f^0; see [17] for a recent study of mode constraints in combination with log-concavity for univariate densities. Retaining notation and assumptions of the previous subsection, we now set

$$F = \Big\{ f \in \text{usc-fcns}(S, [0, \infty]) \ : \ \int f(x)dx = 1, \ C \subset \operatorname{argmax} f, \\ f(x) \leq h(x), \ |f(x) - f(y)| \leq \kappa\|x - y\|_2, \forall x, y \in S \Big\}$$

for some nonempty $C \subset S$. In view of the previous subsection, it suffices to check the argmax constraint. Let $f^n \xrightarrow{h} f$ and $\bar{x} \in C$. Then, $\bar{x} \in \operatorname{argmax} f^n$ for all n. By Proposition 2.3, $\bar{x} \in \operatorname{argmax} f$. Thus, $C \subset \operatorname{argmax} f$ and we have established that F is closed. Suppose that the actual density $f^0 \in F$.

The upper integral bound of Theorem 3.4 holds in view of the previous subsection. The lower integral bound can also be satisfied by using the arguments there, now also paying attention to C when constructing the normal density. We therefore obtain the same conclusions as in the previous subsection with the added benefit that all estimates, regardless of sample size, will have modes that coincides with modes of the actual density in C. For example, if C consists of a single point at which the actual density has a mode, all estimates will also have modes at that point. We note, however, that the formulation applies equally well for multi-modal situations, i.e., when the densities have multiple global maximizers.

4.6 Height of Modes in a Closed Set

We again return to Subsection 4.4 and adopt all its assumptions as well as that $h(x) \leq \gamma/\|x\|_2^p$ for some $\gamma > 0$ and $p > 1$. In addition, let C be a nonempty closed subset of $[0, \infty]$ and set

$$F = \Bigg\{ f \in \text{usc-fcns}(S, [0,\infty]) \ : \ \int f(x)dx = 1, \ \sup f \in C, \\ f(x) \leq h(x), \ |f(x) - f(y)| \leq \kappa\|x - y\|_2, \forall x, y \in S \Bigg\}.$$

To verify that F is closed, it suffices to check the sup condition and, in fact by Proposition 2.3, the tightness of every hypo-convergent sequence. Suppose that $f^0 \in F$. For $f^n \in F$ and $\varepsilon > 0$, the upper bound $f(x) \leq \gamma/\|x\|_2^p$ implies that there exists a compact set $B_\varepsilon \subset S$ such that $f^n(x) \leq \gamma/\|x\|_2^p \leq \varepsilon$ whenever $x \notin B_\varepsilon$. Then, if $\operatorname{argmax} f^n$ is in B_ε, then $\sup_{S\cap B_\varepsilon} f^n = \sup f^n$. If it is not, then $\sup f^n - \varepsilon \leq \varepsilon - \varepsilon = 0 \leq \sup_{S\cap B_\varepsilon} f^n$. Thus, in either case, we have the condition for tightness satisfied. Consequently, by Proposition 2.3, $\sup f^n \to \sup f$ and F is closed because C is closed.

Again, by Theorem 3.4 we obtain strong consistency of $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F} -(1/n)\sum_{j=1}^n \log f(X^j)$ with the added benefit that all estimates, regardless of sample size, have maximum height in the prescribed set C, in which the actual density also has its maximum height.

4.7 Density Values in Intervals

We return to Subsection 4.4 and its assumptions, but now also introduce a lower bounding function $g : S \to [0, \infty)$, which supplements the upper bounding function h to specify density values. The set

$$F = \Bigg\{ f \in \text{usc-fcns}(S, [0,\infty]) : \int f(x)dx = 1, g(x) \leq f(x) \leq h(x), \\ |f(x) - f(y)| \leq \kappa\|x - y\|_2, \forall x, y \in S \Bigg\}$$

is closed by Proposition 4.5 and we again obtain strong consistency provided that F still contains f^0 and a normal density. The latter requirement can be relaxed, but is part of the (simplified) arguments in Subsection 4.4.

4.8 Approximation of Integral

Approximations of the integral constraint might be mandated for computational reasons. We consider a compact $S \subset \mathbb{R}^d$ and let $0 < \alpha \leq \beta < \infty$ and $\kappa \geq 0$. Then,

$$F = \big\{ f \in \text{usc-fcns}(S, [0, \infty]) \ : \ \alpha \leq f(x) \leq \beta, \\ |f(x) - f(y)| \leq \kappa \|x - y\|_2, \forall x, y \in S \big\}$$

is closed by Propositions 4.4 and 4.5. Suppose that the actual density $f^0 \in F^0 = \{f \in F \ : \ \int f(x)dx = 1\}$, which is closed because $f^n \in F \xrightarrow{h} f$ implies pointwise convergence (Proposition 4.4) and the dominated convergence theorem applies. The upper and lower integral bounds in Theorem 3.6 are trivially satisfied. We now introduce approximations of the integral constraint and let

$$F^n = \{f \in F \ : \ |\psi^n(f) - 1| \leq \delta^n\},$$

where $\delta^n \downarrow 0$ and $\psi^n : F \to [0, \infty]$ is an approximating integral mapping with the following properties:

$$\text{For all } n \in \mathbb{N}, \ \psi^n \text{ is continuous and } \sup_{f \in F} \left| \psi^n(f) - \int f(x)dx \right| \leq \eta \gamma^n$$

where $\eta \geq 0$ and $\gamma^n \downarrow 0$. We separate η from γ^n as the estimator only will depend on the rate γ^n and not on the associated constant η, which therefore in practice can remain unknown. For example, in the case of $S = [a, b] \subset \mathbb{R}$, the usual trapezoidal rule with n evaluations of f satisfies the earlier error bound with $\gamma^n = 1/n$. Continuity of ψ^n (in the hypo-topology) follows in view of the pointwise convergence guaranteed by hypo-convergence in this case. For every n, F^n is closed due to the continuity of ψ^n.

Next, we establish that $F^n \to F^0$ under the additional assumption that $\gamma^n / \delta^n \to 0$. If $f \in \operatorname{OutLim} F^n$, then there exists a sequence $f^k \in F^{n_k} \xrightarrow{h} f$. Thus, $|\psi^{n_k}(f^k) - 1| \leq \delta^{n_k}$. Since $\psi^{n_k}(f^k) \to \int f(x)dx$ and $\delta^{n_k} \to 0$, $\int f(x)dx = 1$ and $f \in F^0$. Thus, $\operatorname{OutLim} F^n \subset F^0$. Next, let $f \in F^0$. There exists $\bar{n}$ such that $\eta \gamma^n \leq \delta^n$ for all $n \geq \bar{n}$. For such n, $|\psi^n(f) - 1| \leq \eta \gamma^n \leq \delta^n$ and $f \in F^n$. Consequently, $f \in \operatorname{InnLim} F^n$ and $F^0 \subset \operatorname{InnLim} F^n$. We have established that $F^n \to F^0$ and Theorem 3.6 applies. Hence, $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F^n} -(1/n) \sum_{j=1}^n \log f(X^j)$, which only relies on approximating integral calculations, is strongly consistent provided that $\varepsilon^n \searrow 0$.

4.9 Concavity, Subgradients, Moment Approximations, and Penalization

Next we consider several elements. Suppose that $S \subset \mathbb{R}^d$ is compact with a nonempty interior, $0 < \alpha \leq 1/m(S) \leq \beta < \infty$, with $m(S) = \int_S dx$. For some $\kappa \geq 0$, we set

$$F = \Bigg\{ f \in \text{usc-fcns}(S, [0, \infty]) \ : \ \int f(x)dx = 1, \ f \text{ concave}, \\ \alpha \leq f(x) \leq \beta, \ \|v\|_2 \leq \kappa, \ \forall x \in S, v \in \partial f(x) \Bigg\},$$

which consists of concave densities, bounded from above and away from zero, that also have bounded subgradients. The bound on the subgradients together with the concavity of $f \in F$ ensure that f is Lipschitz continuous with modulus κ. Propositions 4.1, 4.4, and 4.5 together with the dominated convergence theorem ensure that F is closed and obviously nonempty.

Further we introduce first-order moment restrictions and set $F^0 = \{f \in F \ : \ \int xf(x)dx \in C\}$ and $F^n = \{f \in F \ : \ \int xf(x)dx \in C^n\}$, where $C \subset C^n \subset I\!R^d$ are nonempty closed sets and $C^n \to C$. For example, C^n might be some confidence region for the mean values, which at least with high probability contains the actual means. Clearly, both F^0 and F^n are closed by the dominated convergence theorem. Suppose that the actual density $f^0 \in F^0$.

We also consider the penalty function $r^n : F \to [0, \infty)$ given by $r^n(f) = \delta^n \sup f$ for some $\delta^n \downarrow 0$. In view of Proposition 2.3, it is clear that r^n is continuous and also hypo-converge to the zero function. The upper and lower integral bounds in Theorem 3.6 trivially hold in this case and it only remains to show that $F^n \to F^0$. Since $C \subset C^n$, $F^0 \subset F^n$ and it suffices to establish that $\operatorname{OutLim} F^n \subset F^0$. Take $f \in \operatorname{OutLim} F^n$. There exists $f^k \in F^{n_k} \to f$. Since $\int xf^k(x)dx \in C^{n_k}$ and that integral converges to $\int xf(x)dx$, the fact that $C^n \to C$ implies $\int xf(x)dx \in C$. Thus, $f \in F^0$ and $F^n \to F^0$.

An application of Theorem 3.6 then again establishes strong consistency now for a relatively complex estimator $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F^n} -(1/n)\sum_{j=1}^n \log f(X^j) + r^n(f)$ involving multiple shape constraints (bounds, concavity, and subgradient information), penalization that encourages lower modes, and imprecise information about the expected value. Plug-in estimators of modes, near-modes, height of modes, and high-likelihood events are also consistent by a direct application of Theorem 3.8. In this case, tightness of hypo-convergence is automatic because S is compact.

4.10 Epi-Spline Approximations

We next examine the effect of spline approximations and justify their use in implementations; see related developments in [42, 40, 41]. We let $S \subset I\!R^d$ be compact and $0 < \alpha \leq \beta < \infty$, $g : S \to [\alpha, \beta]$, $h \in \text{usc-fcns}(S; [0, \infty])$, with $g(x) \leq h(x) \leq \beta$ for $x \in S$, and $\kappa \geq 0$. The set

$$F = \big\{ f \in \text{usc-fcns}(S, [0, \infty]) \ : \ g(x) \leq f(x) \leq h(x), \\ |f(x) - f(y)| \leq \kappa \|x - y\|_2, \forall x, y \in S \big\}$$

is closed by Propositions 4.4 and 4.5. Suppose that $f^0 \in F^0 = \{f \in F \ : \ \int f(x)dx = 1\}$, which is closed because $f^n \in F \xrightarrow{h} f$ implies pointwise convergence (Proposition 4.4) and the dominated convergence theorem applies. The upper and lower integral bounds in Theorem 3.6 are trivially satisfied.

We utilize *epi-splines* developed in [51] for approximations of lsc functions in the sense of epi-convergence, though we will not need the full generality here and concentrate on simplicial complex partition of S. Thus, S must be polyhedral. Specifically, a *simplex* in $I\!R^d$ is the convex hull of $d + 1$ points $x^0, x^1, ..., x^d \in I\!R^d$, with $x^1 - x^0$, $x^2 - x^0$, ..., $x^d - x^0$ linearly independent. Let the closure of a set $A \subset I\!R^d$ be denoted by $\operatorname{cl} A$. A collection $\mathcal{R} = \{R_k\}_{k=1}^N$ of open subsets of S is a *simplicial complex partition* of S if $\operatorname{cl} R_1$, ..., $\operatorname{cl} R_N$ are simplexes and $\cup_{k=1}^N \operatorname{cl} R_k = S$, and $R_k \cap R_l = \emptyset, k \neq l$. Suppose that

$\{\mathcal{R}^n\}_{n=1}^{\infty}$, with $\mathcal{R}^n = \{R_k^n\}_{k=1}^{N^n}$, is a collection of simplicial complex partition of S with mesh size

$$\mathrm{msh}(\mathcal{R}^n) := \max_{k=1,\ldots,N^n} \sup_{x,y\in R_k^n} \|x-y\|_2 \to 0 \text{ as } n\to\infty.$$

An epi-spline s of order $p \in \mathbb{N}$ on simplicial complex $\mathcal{R} = \{R_k\}_{k=1}^{N}$ is a real-valued function on S that on each R_k is polynomial with total degree p and that satisfies $\liminf s(x^n) = s(x)$ for all $x^n \to x$, i.e., it is lsc. Let e-spl$^p(\mathcal{R}^n)$ be the collection of all such epi-splines; see [51] for details. The approximating set from which estimators will be selected is constructed from such epi-splines as

$$F^n = \left\{ f \in F \ : \ \left| \int f(x)dx - 1 \right| \leq \kappa m(S)\mathrm{msh}(\mathcal{R}^n), \ \ -f \in \text{e-spl}^p(\mathcal{R}^n) \right\},$$

where again $m(S) = \int_S dx$. Relative to F^0, F^n contains two approximations. First, general functions are replaced by epi-splines, or more precisely, by functions $-f \in$ e-spl$^p(\mathcal{R}^n)$. The need for considering the negative is mandated by our present focus on usc densities instead of lsc ones. Second, the integral constraint is relaxed.

Each F^n is closed due to the convergence of integrals under hypo-convergence on F and the fact that the limit of polynomials on a subset of S with values in $[\alpha, \beta]$ must be a polynomial. Moreover, we assume that F^n is nonempty. This is trivially achieved by ensuring that there exists an $-f \in$ e-spl$^p(\mathcal{R}^n)$ "between" g and h. It only remains to show that $F^n \to F^0$. First, we establish that $\mathrm{OutLim}\, F^n \subset F^0$. Let $f \in \mathrm{OutLim}\, F^n$. Then, there exists $f^k \in F^{n_k} \xrightarrow{h} f$. Since $|\int f^k(x)dx - 1| \leq \kappa m(S)\mathrm{msh}(\mathcal{R}^{n_k}) \to 0$ and $\int f^k(x)dx \to \int f(x)dx$, we conclude that $f \in F^0$. Second, we establish that $\mathrm{InnLim}\, F^n \supset F^0$ and let $f \in F^0$. Theorem 3.8 in [51] ensures that there exists continuous $-f^n \in$ e-spl$^1(\mathcal{R}^n)$, with $f^n \xrightarrow{h} f$. An examination of the corresponding proof shows that f^n can be selected to be Lipschitz continuous with modulus κ and $|f^n(x) - f(x)| \leq \kappa \mathrm{msh}(\mathcal{R}^n)$ for $x \in S$. Thus, $|\int f^n(x)dx - 1| \leq \kappa m(S)\mathrm{msh}(\mathcal{R}^n)$ and $f^n \in F^n$, which implies that $f \in \mathrm{InnLim}\, F^n$. We have established that $F^n \to F^0$ and Theorem 3.6 applies. The estimator $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F^n} -(1/n)\sum_{j=1}^n \log f(X^j)$ is therefore strongly consistent provided that $\varepsilon^n \searrow 0$. In view of the construction of F^n, the Estimation Problem in this case involves optimization over $(d+1)N^n$ parameters, where N^n is the number of simplexes used for sample size n (and not N to power n), provided that first-order epi-splines are adopted, i.e., $p = 1$.

4.11 More Possibilities

We end by briefly pointing towards some of the many additional possibilities. Proposition 4.2 shows that restrictions to *s-concave* and other transformations of convex and concave functions by strictly monotone functions result in constraints that are closed subsets of $(\text{usc-fcns}(S;[0,\infty]), d\!\!l)$. Thus, such restrictions naturally fit within our framework and can easily be combined with *any* of the other constraints discussed above. For instance, the existence of estimators under such shape restrictions follows immediately from Theorem 2.6; see [54] for related existence results, but there under growth and smoothness assumptions on the function carrying out the transformation.

The interesting class of *multivariate totally positive densities of order two* (see for example [23]) is also naturally handled by our framework. Recall that such densities have the property that $f(x)f(y) \leq$

$f(\min\{x,y\})f(\max\{x,y\})$ for all $x, y \in S$. Here the min and max are taken componentwise. Under any of the situations discussed above when $f^n \xrightarrow{h} f$ implies pointwise convergence, we have that the restriction to multivariate totally positive densities of order two is a closed set.

In practice, a density estimator is often used as input to a *plug-in estimator*. We already established the consistency of plug-in estimators of modes, near-modes, height of modes, and high-likelihood events, and many others follow similarly. For example, the mean estimator

$$\hat{\mu}^n = \int x \hat{f}^n(x) dx$$

obtained from the density estimator $\hat{f}^n$ is strongly consistent under the assumptions of Subsection 4.9 and in fact under more relaxed assumptions too.

Acknowledgements. This work is supported in part by the CIMS program at the Naval Postgraduate School and ONR Science of Autonomy Program under N0001417WX01210 and N00014-17-1-2372.

A Proofs

Proof of Proposition 2.3: For part (i), we first suppose that $\sup f$ is finite and let $\varepsilon > 0$. There exists $x \in S$ such that $f(x) \geq \sup f - \varepsilon$ and also by the definition of hypo-convergence, $x^n \in S \to x$ such that $\liminf f^n(x^n) \geq f(x)$. Thus, $\liminf(\sup f^n) \geq \liminf f^n(x^n) \geq f(x) \geq \sup f - \varepsilon$. Second, suppose that $\sup f = \infty$ and let $\delta > 0$. Then, there exists $x \in S$ such that $f(x) \geq \delta$ and also by the definition of hypo-convergence, $x^n \in S \to x$ such that $\liminf f^n(x^n) \geq f(x)$. Thus, $\liminf(\sup f^n) \geq \liminf f^n(x^n) \geq f(x) \geq \delta$. Since ε and δ are arbitrary, part (i) is established.

For part (ii), suppose $\bar{x} \in \operatorname{OutLim}(\delta^n\text{-}\operatorname{argmax} f^n)$. Then, there exists $\{x^k \in \delta^{n_k}\text{-}\operatorname{argmax} f^{n_k},\ k \in I\!N\} \to \bar{x}$. Thus, $\liminf_k f^{n_k}(x^k) \geq \liminf_k (\sup f^{n_k} - \delta^{n_k}) \geq \sup f - \delta$ by part (i). In view of the definition of hypo-convergence, this implies that $f(\bar{x}) \geq \limsup_k f^{n_k}(x^k) \geq \liminf_k f^{n_k}(x^k) \geq \sup f - \delta$. Hence, $\bar{x} \in \delta\text{-}\operatorname{argmax} f$ and part (ii) is established.

Part (iii) is immediate from the fact that if $\bar{x} \in \operatorname{OutLim}(\operatorname{lev}_\alpha f^n)$, then there exists $x^k \in \operatorname{lev}_\alpha f^{n_k} \to \bar{x}$ and thus by the definition of hypo-convergence, $f(\bar{x}) \geq \limsup f^{n_k}(x^k) \geq \alpha$.

For part (iv), we observe in view of parts (i) and (ii) that $\bar{x} \in \operatorname{argmax} f$ and also $\liminf(\sup f^n) \geq \sup f$. The definition of hypo-convergence implies that $\limsup_k(\sup f^{n_k}) = \limsup_k f^{n_k}(x^{n_k}) \leq f(\bar{x}) = \sup f$ and the conclusion holds. Part (v) holds by [47, Theorem 7.31]. □

Proof of Proposition 4.1: For part (i), we recall that $\operatorname{hypo} f^n \to \operatorname{hypo} f$ for concave f^n implies that $\operatorname{hypo} f$ is convex and thus f is concave [47, Proposition 4.15]. The implication for uniform convergence is a consequence of [47, Theorem 7.17]. Since $-f^n, -f$ are proper, lsc, and convex, it follows by [47, Theorem 12.35] that the graphs of the subdifferentials ∂f^n set-converge to the graph of ∂f. Thus, for every (x, v) in the graph of ∂f, there exists $x^n \to x$ and $v^n \to v$, with $v^n \in \partial f^n(x^n)$. Since $\|v^n\|_2 \leq \kappa$ for all n, we also have that $\|v\|_2 \leq \kappa$, which establishes the claim.

For part (ii), we recall the extension of the log-function to $[0, \infty]$ and that it is strictly increasing and continuous. The compositions $\log f^n$ and $\log f$ are usc, and also $\log f^n \xrightarrow{h} \log f$ as established next.

Suppose that $x^n \in S \to x$. Then, $\operatorname{limsup} f^n(x^n) \leq f(x)$ and therefore also $\operatorname{limsup} \log f^n(x^n) \leq \log f(x)$. Similarly, for all $x \in S$, there exists $x^n \in S \to x$ with $\operatorname{liminf}(\log f^n(x^n)) \geq \log f(x)$. Thus, $\log f^n \xrightarrow{h} \log f$. Since $\log f^n$ is concave, we much have that $\log f$ is concave too; again invoking Proposition 4.1. Thus, f is log-concave. By [47, Theorem 7.17], we have that for every compact set $C \subset \operatorname{int}\{x \in S \ : \ \log f(x) > -\infty\}$, we have that $\log f^n$ converges to $\log f$ uniformly on C. The last part of (ii) is then an algebraically obtained restatement of this uniform convergence.

For (iii), let $\lambda \in (0,1)$ and $x, y \in \operatorname{int} S$. Set $z = \lambda x + (1-\lambda)y$. Hypo-convergence implies that there exists $z^n \in \operatorname{int} S \to z$ such that $f^n(z^n) \to f(z)$. Construct $x^n = x + z^n - z$ and $y^n = y + z^n - z$. Clearly, $x^n \to x$ and $y^n \to y$. Then, $\lambda x^n + (1 - \lambda y^n) = z^n$. Let $\varepsilon > 0$. There exists $\bar{n}$ such that for all $n \geq \bar{n}$, $x^n, y^n \in S$ and

$$f(z) \leq f^n(z^n) + \frac{\varepsilon}{3}, \quad f^n(x^n) \leq f(x) + \frac{\varepsilon}{3\lambda}, \quad f^n(y^n) \leq f(y) + \frac{\varepsilon}{3(1-\lambda)}.$$

Collecting these results and use the convexity of f^n, we obtain that for $n \geq \bar{n}$

$$f(z) \leq f^n(z^n) + \frac{\varepsilon}{3} \leq \lambda f^n(x^n) + (1-\lambda) f^n(y^n) + \frac{\varepsilon}{3} \leq \lambda f(x) + (1-\lambda) f(y) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, $f(z) \leq \lambda f(x) + (1-\lambda) f(y)$. It only remains to examine the case when x and/or y are at the boundary of S. Suppose that $\lambda \in (0,1)$, $x \in \operatorname{int} S$, and $y \in S \setminus \operatorname{int} S$. Then, there exists $y^n \in \operatorname{int} S \to y$ with $f(\lambda x + (1-\lambda)y^n) \leq \lambda f(x) + (1-\lambda) f(y^n)$. Since $\lambda x + (1-\lambda)y^n, \lambda x + (1-\lambda)y \in \operatorname{int} S$ and f is continuous on $\operatorname{int} S$, the left-hand side tends to $f(\lambda x + (1-\lambda)y)$. The upper limit of the right-hand side is $\lambda f(x) + (1-\lambda) f(y)$ in view of the usc of f. A similar argument holds in the other cases. Thus, f is convex. □

Proof of Proposition 4.2: The inverse h^{-1} exists and is strictly increasing and continuous. Thus, $f^n = h(g^n(\cdot))$ implies that $g^n = h^{-1}(f^n(\cdot))$. It follows directly from the definition of hypo-convergence that $g^n \xrightarrow{h} g = h^{-1}(f(\cdot))$. For part (i), g must be concave in view of part (i) of Proposition 4.1 and $f = h(g(\cdot))$. Part (ii) follows by a similar argument but now invoking part (iii) of Proposition 4.1. □

Proof of Proposition 4.3: For part (i), let $x \leq y$, with $y \in \operatorname{int} S$, and $\varepsilon > 0$. The usc property implies that there exists $\delta > 0$ such that $f(y) \geq f(z) - \varepsilon$ for all $z \in S$ with $\|z - y\|_\infty \leq \delta$. Since $y \in \operatorname{int} S$, z can be takes such that $z_i > y_i$ for $i = 1, ..., d$ and $z \in \operatorname{int} S$. In view of the hypo-convergence, there exists $x^n \in S \to x$ such that $f(x) \leq \operatorname{liminf} f^n(x^n)$ and also $\operatorname{limsup} f^n(z) \leq f(z)$. Thus, $x^n \leq z$ for sufficiently large n. Using the nondecreasing property, we then obtain that

$$f(x) \leq \operatorname{liminf} f^n(x^n) \leq \operatorname{liminf} f^n(z) \leq \operatorname{limsup} f^n(z) \leq f(z) \leq f(y) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary the first conclusion follows.

Under the additional structure of S, the argument can be modified as follows. Now with $y \in S$, let $\delta > 0$ and x^n be as earlier. Construct $z \in I\!R^d$ be setting $z_i = \min\{\beta_i, y_i + \delta\}$. Let $\bar{n}$ be such that $x_i^n \leq x_i + \delta$ for all $i = 1, ..., d$ and $n \geq \bar{n}$. Then, for $n \geq \bar{n}$, $x_i^n \leq \min\{\beta_i, x_i + \delta\} \leq \min\{\beta_i, y_i + \delta\} = z_i$. Thus, again we have that $x^n \leq z$ for sufficiently large n and the preceding arguments lead to the conclusion.

For (ii) let $x \leq y$, with $x \in \operatorname{int} S$, and $\varepsilon > 0$. The usc property implies that there exists $\delta > 0$ such that $f(x) \geq f(z) - \varepsilon$ for all $z \in S$ with $\|z - x\|_\infty \leq \delta$. Since $x \in \operatorname{int} S$, z can be takes such that $z_i < x_i$ for $i = 1, ..., d$ and $z \in \operatorname{int} S$. In view of the hypo-convergence, there exists $y^n \in S \to y$ such that $f(y) \leq \operatorname{liminf} f^n(y^n)$ and also $\operatorname{limsup} f^n(z) \leq f(z)$. Thus, $z \leq y^n$ for sufficiently large n. Using the nonincreasing property, we then obtain that

$$f(y) \leq \operatorname{liminf} f^n(y^n) \leq \operatorname{liminf} f^n(z) \leq \operatorname{limsup} f^n(z) \leq f(z) \leq f(x) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary the first conclusion follows.

Under the additional structure of S, the argument can be modified as follows. Now with $x \in S$, let $\delta > 0$ and y^n be as earlier. Construct $z \in I\!R^d$ be setting $z_i = \max\{\alpha_i, x_i - \delta\}$. Let $\bar{n}$ be such that $y_i^n \geq y_i - \delta$ for all $i = 1, ..., d$ and $n \geq \bar{n}$. Then, for $n \geq \bar{n}$, $y_i^n \geq \max\{\alpha_i, y_i - \delta\} \geq \max\{\alpha_i, x_i - \delta\} = z_i$. Again we have $z \leq y^n$ for sufficiently large n and the preceding arguments lead to the conclusion. □

Proof of Proposition 4.4: If $\kappa = 0$, then f^n are constant functions on S and f also, and the conclusion holds. Suppose that $\kappa > 0$. Let $x, y \in S$ and $\varepsilon > 0$. Hypo-convergence implies that there exists $x^n \in S \to x$ such that $f^n(x^n) \to f(x)$ and $\operatorname{limsup} f^n(y) \leq f(y)$. Hence, there exists $\bar{n}$ such that for all $n \geq \bar{n}$, $\|x^n - x\| \leq \varepsilon/(3\kappa)$, $|f^n(x^n) - f(x)| \leq \varepsilon/3$, $f^n(y) \leq f(y) + \varepsilon/3$. For such n, $f(x) - f(y)$

$$\begin{aligned} &= f(x) - f^n(x^n) + f^n(x^n) - f^n(x) + f^n(x) - f^n(y) + f^n(y) - f(y) \\ &\leq \frac{\varepsilon}{3} + \kappa\|x^n - x\| + \kappa\|x - y\| + f(y) + \frac{\varepsilon}{3} - f(y) \leq \kappa\|x - y\| + \varepsilon. \end{aligned}$$

Repeating this argument with the roles of x and y interchanged, we obtain that $|f(x) - f(y)| \leq \kappa\|x - y\| + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, f is Lipschitz continuous with modulus κ.

To establish the uniform convergence, let $\varepsilon > 0$ and $\{z^j\}_{j=1}^m \subset C$ be such that $\cup_{j=1}^m I\!B(z^j, \varepsilon/(4\kappa)) \supset C$, where $I\!B(x, r) := \{x' \in S : \|x' - x\| \leq r\}$. For each j, there exist $\bar{n}^j$ and $z^{nj} \in S \to z^j$ such that $|f^n(z^{nj}) - f(z^j)| \leq \varepsilon/4$ and $\|z^{nj} - z^j\| \leq \varepsilon/(4\kappa)$ for all $n \geq \bar{n}^j$. Set $\bar{n} = \max\{\bar{n}^1, ..., \bar{n}^m\}$. For $n \geq \bar{n}$ and $x \in I\!B(z^j, \varepsilon/(4\kappa))$, $|f(x) - f^n(x)|$

$$\begin{aligned} &\leq |f(x) - f(z^j)| + |f(z^j) - f^n(z^{nj})| + |f^n(z^{nj}) - f^n(z^j)| + |f^n(z^j) - f^n(x)| \\ &\leq \kappa\frac{\varepsilon}{4\kappa} + \frac{\varepsilon}{4} + \kappa\frac{\varepsilon}{4\kappa} + \kappa\frac{\varepsilon}{4\kappa} = \varepsilon. \end{aligned}$$

Since the same result holds for all j, the conclusion follows. □

Proof of Proposition 4.5: Let $x \in S$ and observe that $g(x) \leq \operatorname{limsup} f^n(x) \leq f(x)$, which established the lower bound. Since h is usc, we also have that for some $x^n \in S \to x$, $h(x) \geq \operatorname{limsup} h(x^n) \geq \operatorname{liminf} f^n(x^n) \geq f(x)$, which confirms the upper bound. □

References

[1] C. Choirat abd C. Hess. A functional verion of the birkoff ergodic theorem for a normal integrand: a variational approach. *Annals of Probability*, 31:63–92, 2003.

[2] Z. Artstein and R. J-B Wets. Consistency of mimnimizers and the SLLN for stochastic programs. *J. Convex Analysis*, 2:1–17, 1996.

[3] F. Balabdaoui, H. Jankowski, M. Pavlides, A. Seregin, and J. Wellner. On the Grenander estimator at zero. *Statistica Sinica*, 21(2):873–899, 2011.

[4] F. Balabdaoui, K. Rufiback, and J. A. Wellner. Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of Statistics*, 37:1299–1331, 2009.

[5] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: limit distribution theory and the spline connection. *Annals of Statistics*, 35(6):2536–2564, 2007.

[6] F. Balabdaoui and J. A. Wellner. Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

[7] L. Birgé. Estimation of unimodal densities without smoothness assumptions. *Annals of Statistics*, 25:970–981, 1997.

[8] M. Birke. Shape constrained kernel density estimation. *J. Statistical Planning and Inference*, 139:2851–2862, 2009.

[9] X. Chen. Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometric*, pages 5549–5632. 2007. Volume 6B, Chapter 76.

[10] M. Cule, R.J. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *J. Royal Statistical Society Series B*, 72:545–600, 2010.

[11] M. Cule, R.J. Samworth, and M. Stewart. Rejoinder to maximum likelihood estimation of a multi-dimensional log-concave density. *J. Royal Statistical Society Series B*, 72:600–607, 2010.

[12] M. L. Cule and R. J. Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic J. Statistics*, 4:254–270, 2010.

[13] L. Dechevsky and S. Penev. On shape-preserving probabilistic wavelet approximators. *Stochastic Analysis and Applications*, 15:187–215, 1997.

[14] R.A. DeVore. Monotone approximation by polynomials. *SIAM J. Mathematical Analysis*, 8:906–921, 1977.

[15] R.A. DeVore. Monotone approximation by splines. *SIAM J. Mathematical Analysis*, 8:891–905, 1977.

[16] M. X. Dong and R. J-B Wets. Estimating density functions: a constrained maximum likelihood approach. *J. Nonparametric Statistics*, 12(4):549–595, 2000.

[17] C. R. Doss and J. A. Wellner. Mode-constrained estimation of a log-concave density. *ArXiv e-prints*, November 2016.

[18] C.R. Doss. *Shape-Constrained Inference for Concave-Transformed Densities and their Modes*. PhD dissertation, University of Washington, 2013.

[19] L. Dumbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68, 2009.

[20] J. Dupacova. Epi-consistency in restricted regression models - the case of a general convex fitting function. *Computational Statistics and Data Analysis*, 14:417–425, 1992.

[21] R. A. Durrett. *Probability : Theory and Examples*. Duxbury Press, 2. edition, 1996.

[22] P.B. Eggermont and V.N. LaRiccia. *Maximum Penalized Likelihood Estimation, Volume I: Density Estimation*. Springer, 2001.

[23] S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in markov structures. *Annals of Statistics*, 2017.

[24] F. Gao and J. A. Wellner. On the rate of convergence of the maximum likelihood estimator of a k-monotone density. *Science in China Series A: Mathematics*, 52(7), 2009.

[25] S. Geman and C.-R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10(2):401–414, 1982.

[26] U. Grenander. On the theory of mortality measurement. II. *Skandinavisk Aktuarietidskrift*, 39:125–153, 1956.

[27] U. Grenander. *Abstract Inference*. Wiley, 1981.

[28] P. Groeneboom and G. Jongbloed. *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge University Press, 2014.

[29] P. Groeneboom, G. Jongbloed, and J.A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics*, 29(6):1653–1698, 2001.

[30] P. Groeneboom and J.A. Wellner. *Information bounds and nonparametric maximum likelihood estimation*. Birkhauser, Basel, 1992.

[31] P. Hall and K.-H. Kang. Unimodal kernel density estimation by data sharpening. *Statistica Sinica*, 15:73–98, 2005.

[32] G. Jongbloed. The iterative convex minorant algorithm for nonparametric estimation. *J. Computational and Graphical Statistics*, 7:310–321, 1998.

[33] A. K. H. Kim, A. Guntuboyina, and R. J. Samworth. Adaptation in log-concave density estimation. *ArXiv e-prints*, September 2016.

[34] A. K. H. Kim and R. J. Samworth. Global rates of convergence in log-concave density estimation. *Annals of Statistics*, 44:2756–2779, 2016.

[35] V. K. Klonias. Consistency of two nonparametric maximum penalized likelihood estimators of the probability density function. *Annals of Statistics*, 10:811–824, 1982.

[36] R. Koenker and I. Mizera. Density estimation by total variation regularization. In *A Festschrift for Kjell Doksum*. World Scientific, Singapore, 2006.

[37] R. Koenker and I. Mizera. Quasi-concave density estimation. *Annals of Statistics*, 38:2998–3027, 2010.

[38] L. A. Korf and R. J-B Wets. Random lsc functions: an ergodic theorem. *Mathematics of Operations Research*, 26(2):421–445, 2001.

[39] T. Leonard. Density estimation, stochastic processes and prior information. *J. Royal Statistical Society*, B40:113–146, 1978.

[40] M. Meyer. Constrained penalized splines. *Canadian J. Satistics*, 40:190–206, 2012.

[41] M. Meyer. Nonparametric estimation of a smooth density with shape restrictions. *Statistica Sinica*, 22:681–701, 2012.

[42] M. Meyer and D. Habtzghib. Nonparametric estimation of density and hazard rate functions with shape restrictions. *J. Nonparametric Statistics*, 23(2):455–470, 2011.

[43] J. Kumar Pal, M. Woodroofe, and M. Meyer. Estimating a polya frequency function. In R. Liu, W. Strawderman, and C.-H. Zhang, editors, *Complex datasets and inverse problems*, pages 239–249. Beachwood, OH, 2007. IMS Lecture Notes Monogr. Ser., volume 54.

[44] D. Papp. *Estimation problems involving nonnegative polynomials and their restrictions*. PhD dissertation, Rutgers University, 2011.

[45] D. Papp and F. Alizadeh. Shape constrained estimations using nonnegative splines. *J. Computational and Graphical Statistics*, 23(1):211–231, 2014.

[46] L. Reboul. Estimation of a function under shape restrictions. applications to reliability. *Annals of Statistics*, 33:1330–1356, 2005.

[47] R.T. Rockafellar and R. J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaft*. Springer, 3rd printing-2009 edition, 1998.

[48] J. O. Royset. Approximations and solution estimates in optimization. *Mathematical Programming*, To appear, 2017. Preprint at http://faculty.nps.edu/joroyset/pubs.html.

[49] J. O. Royset and R. J-B Wets. Lopsided convergence: an extension and its quantifications. In review. Preprint at http://faculty.nps.edu/joroyset/pubs.html.

[50] J. O. Royset and R. J-B Wets. Fusion of hard and soft information in nonparametric density estimation. *European J. Operational Research*, 247(2):532–547, 2015.

[51] J. O. Royset and R. J-B Wets. Multivariate epi-splines and evolving function identification problems. *Set-Valued and Variational Analysis*, 24(4):517–545, 2016. Erratum: pp. 547-549.

[52] K. Rufiback. Computing maximum likelihood estimators of a log-concave density function. *J. Statistical Computation and Simulation*, 77:561–574, 2007.

[53] E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression. *Annals of Statistics*, 39:1633–1657, 2011.

[54] A. Seregin and J. A. Wellner. Nonparametric estimation of multivariate convex-transformed densities. *Annals of Statistics*, 38(6):3751–3781, 2010.

[55] B. W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10:795–810, 1982.

[56] J. R. Thompson and R. A. Tapia. *Nonparametric Function Estimation, Modeling, and Simulation*. SIAM Publishers, Philadelphia, PA, 1990.

[57] A. W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 2nd printing 2000 edition, 1996.

[58] M. Valadier. Conditional expectation and ergodic theorem for a positive integrand + corrigendum. *J. Nonlinear Convex Analysis*, 1, 3:233–244, 123, 2000, 2002.

[59] A. W. van der Vaart. Empirical processes and statistical learning. Lecture Notes, Vrije Universiteit, Amsterdam, Netherland, 2011.

[60] G. Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

[61] J. Wang. Asymptotics of least-squares estimators for constrained nonlinear regression. *Annals of Statistics*, 24(3):1316–1326, 1996.

[62] M. Woodroofe and J. Sun. A penalized maximum likelihood estimate $f(0+)$ when f is non-decreasing. *Statistica Sinica*, 3:501–515, 1993.