2019

Variational Analysis of Constrained M-Estimators

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

ArXiv

http://hdl.handle.net/10945/64720

Variational Analysis of Constrained M-Estimators

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 unified framework for establishing existence of nonparametric M-estimators, computing the corresponding estimates, and proving their strong consistency when the class of functions is exceptionally rich. In particular, the framework addresses situations where the class of functions is complex involving information and assumptions about shape, pointwise bounds, location of modes, height at modes, location of level-sets, values of moments, size of subgradients, continuity, distance to a "prior" function, multivariate total positivity, and any combination of the above. The class might be engineered to perform well in a specific setting even in the presence of little data. The framework views the class of functions as a subset of a particular metric space of upper semicontinuous functions under the Attouch-Wets distance. In addition to allowing a systematic treatment of numerous M-estimators, the framework yields consistency of plug-in estimators of modes of densities, maximizers of regression functions, level-sets of classifiers, and related quantities, and also enables computation by means of approximating parametric classes. We establish consistency through a one-sided law of large numbers, here extended to sieves, that relaxes assumptions of uniform laws, while ensuring global approximations even under model misspecification.

Keywords: shape-constrained estimation, variational approximations

Date: September 11, 2019

1 Introduction

It is apparent that the class of functions from which nonparametric M-estimators are selected should incorporate non-data information about the stochastic phenomenon under consideration and also modeling assumptions the statistician would like to explore. In applications, the class can become complex involving shape restrictions, bounds on moments, slopes, modes, and supports, limits on tail characteristics, constraints on the distance to a "prior" distribution, and so on. The class might be engineered to perform well in a particular setting; statistical learning

is often carried out with highly engineered estimators. An ability to consider rich classes of functions leads to novel estimators that even in the presence of relatively little data can produce reasonable results.

Numerous theoretical and practical challenges arise when considering M-estimators selected from rich classes of functions on $\mathbb{R}^d$, which may even be misspecified, as we need to analyze and solve infinite-dimensional random optimization problems with nontrivial constraints. In this article, we leverage and extend results from Variational Analysis to build a unified framework for establishing existence of such *constrained M-estimators*, computing the corresponding estimates, and proving their strong consistency. We also show strong consistency of plug-in estimators of modes of densities, maximizers of regression functions, level-sets of classifiers, and related quantities that likewise account for a variety of constraints. In contrast to "classical" analysis, Variational Analysis centers on functions that abruptly change due to constraints and other sources of nonsmoothness and therefore emerges as a natural tool for examining M-estimators selected from rich classes of functions.

1.1 Setting and Challenges

Given d_0-dimensional random vectors X^1, X^2, ..., X^n, we consider constrained M-estimators of the form

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

where F^n is a class of candidate functions on $\mathbb{R}^d$, or a subset thereof, possibly varying with n (sieved), ψ is a loss function such as $\psi(x, f) = -\log f(x)$ (maximum likelihood (ML) estimation of densities) and $\psi((x, y), f) = (y-f(x))^2$ (least-squares (LS) regression), π^n is a penalty function possibly introduced for the purpose of smoothing and regularization, and the inclusion of $\varepsilon^n \geq 0$ indicates that near-minimizers are permitted. We focus on the iid case, but extensions to non-iid samples is possible within our framework.

The Grenander estimator, the ML estimator over log-concave densities, and the LS regression function under convexity, just to mention a few constrained M-estimators, certainly exist. However, existence is not automatic. For rich classes of functions, it is rather common to have an empty set of minimizers in (1); Section 2 furnishes examples. The extensive literature on M-estimators establishes consistency under rather general conditions (see, e.g., [44, Thm. 3.2.2, Cor. 3.2.3], [46, Thm. 5.7], and [45, Thms. 4.3, 4.8]). Standard arguments pass through uniform convergence of $n^{-1}\sum_{j=1}^{n} \psi(X^j, \cdot)$ to $\mathbb{E}[\psi(X^1, \cdot)]$, almost surely or in probability, on a sufficiently large class of functions, which in turn reduces to checking integrability and total boundedness of the class under an appropriate (pseudo-)metric, the latter being equivalent to finite metric entropy. It has long been recognized that uniform convergence is unnecessarily strong; already Wald [49] adopted a weaker one-sided condition. In the central case of ML estimation of densities, an upper bound on $\psi(x, f) = -\log f(x)$ may not be available and typically force reformulations

in terms of $\psi(x, f) = -\log(f(x) + f^0(x))/2f^0(x)$ and similar expressions, where f^0 is some reference density. Uniform convergence also gives rise to measurability issues, which may require statements in terms of outer measures [44].

In the presence of rich classes of functions, it becomes nontrivial to compute estimates as there are no general algorithm for (1). Approximations in terms of basis functions are not easily constructed because the class of functions may neither be a linear space nor a convex set.

1.2 Contributions

In this article, we address the challenges of existence, consistency, and computations of constrained M-estimators by viewing the class of functions under consideration as a subset of a particular metric space of upper-semicontinuous (usc) functions equipped with the Attouch-Wets (aw)-distance[1]. Although viewing M-estimators as minimizers of empirical processes indexed by a metric space is standard, our *particular* choice is novel. The only precursors are [32, 34], which hint to developments in this direction without a systematic treatment. Three main advantages emerge from the choice of metric space: (i) A unified and disciplined approach to rich classes of functions becomes possible as the aw-distance can be used across M-estimators. (ii) Consistency of plug-in estimators of modes of densities, maximizers of regression functions, level-sets of classifiers, and related quantities follows immediately from consistency of the underlying estimators. (iii) Computation of estimates becomes viable because usc functions, even when defined on unbounded sets, can be approximated by certain parametric classes to an arbitrary level of accuracy in the aw-distance. Moreover, the unified treatment of rich classes of functions allows for a majority of algorithmic components to be transferred from one M-estimator to another.

We bypass uniform laws of large numbers (LLN) and accompanying metric entropy calculations, and instead rely on a one-sided *lsc-LLN* for which upper bounds on the loss function ψ becomes superfluous. Thus, concern about density values near zero and the need for reformulations in ML estimation vanish. Challenges related to measurability reduces to simple checks on the loss function that can be stated in elementary terms. Already Wald [49] and Huber [21] recognized the one-sided nature of (1) and this perspective was subsequently formalized and refined under the name *epi-convergence*; see [15, 50, 33] for results in the parametric case and also [29, Ch. 7]. In the nonparametric case, the use of epi-convergence to establish consistency of M-estimators appears to be limited to [12], which considers ML estimators of densities that are selected from 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. Sieves are not permitted. The Hilbert space setting is problematic as one cannot rely on (strong) compactness to ensure existence of estimators and their cluster points, and weak compactness essentially limits the scope to convex classes of functions. In addition to going much beyond ML estimation, our particular choice

[1]The aw-distance quantifies distances between sets, in this case hypographs (also called subgraphs) and the name hypo-distance is sometimes used; see Sec. 3.

of metric space addresses issues about existence. We also provide a novel consistency result that extends the reach of the lsc-LLN to sieves, which is of independent interest in optimization theory.

Without insisting on uniform approximations, the lsc-LLN establishes convergence in some sense across the whole class of functions. Thus, consistency results are not hampered by model misspecification or other circumstances under which an estimator is constrained away from an actual (true) function. They only need to be interpreted appropriately, for example in terms of minimization of Kullback-Leibler divergence. It also becomes immaterial whether the estimator and the actual function are unique. Under misspecification in ML estimation, just to mention one case, there can easily be an uncountable number of densities that have the same Kullback-Leibler divergence to the one from which the data is generated. Our results still hold.

We construct an algorithm for (1) that under moderate assumptions produces an estimate in a *finite* number of iterations if $\varepsilon^n > 0$ and to converge to an estimate otherwise. The algorithm permits the use of a wide variety of state-of-the-art optimization subroutines. We demonstrate the framework in a small study of ML estimation over densities on $[0,1]^2$ that satisfy pointwise upper and lower bounds, have nonunique modes covering two specific points, are Lipschitz continuous, and are subject to smoothing penalties.

In our framework, conditions for existence and consistency of estimators essentially reduce to checking that the class of functions F^n is closed under the aw-distance. It is well known that the class of concave densities is closed in this sense. We establish that many other natural classes of functions are also closed in the aw-distance. Specifically, we show this for classes defined by convexity, log-concavity, monotonicity, s-concavity, monotone transformations, Lipschitz continuity, pointwise upper and lower bounds, location of modes, height at modes, location of level-sets, values of moments, size of subgradients, splines, multivariate total positivity of order two, and *any combination* of the above, possibly under additional assumptions. To the best of our knowledge, no prior study has established existence and consistency of M-estimators for such a variety of constraints.

We defer the systematic treatment of rates of convergence for M-estimators within the proposed framework. Still, because covering numbers of bounded subsets of usc functions under the aw-distance are known [31], it is immediately clear that under certain (strong) assumptions rate results can be obtained (see [31] for preliminary examples), but these are presently not as sharp as those available by means of empirical process theory.

Section 2 provides motivating examples and a small empirical study. Main results follow in Section 3. Section 4 establishes the closedness of a variety of function classes under the aw-distance. Section 5 states an algorithm for (1) and Section 6 gives additional examples. The paper ends with intermediate results and proofs in Section 7.

2 Motivation and Examples

The study is motivated, in part, by estimation in the presence of relatively little data. In such contexts, constraints in the form of well-selected classes of functions over which to optimize may become useful. Although statistical models often aspire to be tuning-free (see for example [7]), models in statistical learning and related application areas are far from being free of tuning [28]. We follow that recent trend by considering novel nonparametric estimators defined by complex constraints, many of which might be tuned to address specific settings.

2.1 Role of Constraints

Analysis using integral-type metrics such as those defined by L^2 and Hellinger distances leads to many of the well-known results for LS regression and ML estimation of densities. However, difficulties arise with the introduction of constraints, especially related to closedness and compactness of the class of function under consideration. For example, consider the class of bi-constant densities on $[0, 1]$, with each density having one value on $[0, 1/2]$ and potentially another value on $(1/2, 1]$, that also must satisfy $f(x) < 3/2$ for all $x \in [0, 1]$. When the number of samples in $[0, 1/2]$ is sufficiently different from that in $(1/2, 1]$, the ML estimator over this class does not exist as the value of the density in the interval with the more samples would be pushed up towards the unattainable upper bound. The break-down is caused by a class of densities that is not closed. Although rather obvious here, the situation becomes nontrivial in nonparametric cases involving rich classes of functions that may even be misspecified. In fact, already the ML estimators over unimodal densities on $\mathbb{R}$ [3] and over log-concave densities on $\mathbb{R}^d$ for $n \leq d$ [14] fail to exist.

For another example, suppose that the definition of a class includes the constraint that the maximizers of the functions should contain a given point in $\mathbb{R}^d$. This constraint conveys information or assumption about the location of modes in the density setting and "peaks" in a regression problem. A sequence of estimates satisfying this constraint may have L^2 and Hellinger limits that violate it; the constraint is not closed under these metrics. Even the simple constraint that $f(\bar{x}) \geq 1$ for a given $\bar{x} \in \mathbb{R}^d$, which is a constraint on a level-set of f, would not be closed. However, the constraints on maximizers and such level-sets are indeed closed in the aw-distance; see Section 2.2 and, more comprehensively, Section 3.4.

Constraints related to maximizers, maxima, and level-sets motivate the choice of the aw-distance in a profound way as neither pointwise nor uniform convergence would be satisfactory with regard to those: Pointwise convergence fails to ensure convergence of maximizers and uniform convergence applies essentially only to continuous functions defined on compact sets.

2.2 Example Formulation and Result

As a concrete example of a rich class of densities in ML estimation on $\mathbb{R}^d$, suppose that $\alpha, \kappa \geq 0$; $C, D \subset \mathbb{R}^d$; $I \subset [0,\infty]$ is closed; $g, h : \mathbb{R}^d \to [0,\infty)$, with h being usc and also satisfying $\int h(x)dx < \infty$; and

$$\begin{aligned} F = \Big\{ f : \mathbb{R}^d \to [0,\infty] \;\Big|\; & f \text{ usc}, \;\; \int f(x)dx = 1, \\ & C \subset \operatorname{argmax}_{x \in R^d} f(x), \;\; D \subset \operatorname{lev}_{\geq \alpha} f, \;\; \sup_{x \in R^d} f(x) \in I, \\ & g(x) \leq f(x) \leq h(x), \;\; |f(x) - f(y)| \leq \kappa \|x - y\|_2, \forall x, y \in \mathbb{R}^d \Big\}, \end{aligned} \tag{2}$$

where $\operatorname{lev}_{\geq \alpha} f = \{x \in \mathbb{R}^d \mid f(x) \geq \alpha\}$ is an upper level-set of f. The second line restricts the consideration to densities with (global) modes covering C and "high-probability regions" covering D. Neither C nor D need to be singletons. Although there are some efforts towards accounting for information about the location of modes (see for example [13]), the generality of these constraints is unprecedented. The third line permits nearly arbitrary pointwise bounds. In settings with little data but substantial experience about what an estimate "should" look like, such constraints can be helpful modeling tools. The last constraint restricts the class to Lipschitz continuous functions with modulus κ.

Properties of the ML estimator on this class is stated next. Section 7 furnishes the proof and those of most subsequent results. Let $\mathbb{N} = \{1, 2, \dots\}$.

2.1 Proposition *Suppose that $X^1, X^2, \dots$ are iid random vectors, each distributed according to a density $f^0 : \mathbb{R}^d \to [0,\infty]$, F in (2) is nonempty, and $\{\varepsilon^n \geq 0, n \in \mathbb{N}\} \to 0$. Then the following hold almost surely:*

(i) For all $n \in \mathbb{N}$, there exists $\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F}\{-n^{-1}\sum_{j=1}^n \log f(X^j)\}$.

(ii) Every cluster point (under the aw-distance) of $\{\hat{f}^n, n \in \mathbb{N}\}$, of which there is at least one, minimizes the Kullback-Leibler divergence to f^0 over the class F.

(iii) If $f^0 \in F$, then f^0 is the limit (under the aw-distance) of $\{\hat{f}^n, n \in \mathbb{N}\}$.

The proposition establishes that despite the rather rich class, ML estimators exists and they are consistent, even under model misspecification, which is especially relevant in the presence of many constraints. The specific case examined here is only an illustration. Section 3 provides general results along these lines. The framework aims to simplify the analysis of *new* estimators constructed by adding and/or removing constraints. As we see below, the analysis reduces largely to checking closedness and nonemptiness.

2.3 Empirical Results

We consider ML estimation of the mixture of three uniform densities on $[0,1]^2$ depicted in Figure 1(left). The resulting mixture density f^0 has height $f^0(x) = 3$ for x in the areas colored yellow and $f^0(x) = 0.6150$ elsewhere. Using a sample of size 100 shown in Figure 1(right), we compute a penalized ML estimate over the class of functions

$$F = \Big\{ f : [0,1]^2 \to [\alpha, \beta] \;\Big|\; \int f(x)dx = 1,\ \{\bar{x}, \bar{y}\} \subset \operatorname{argmax}_{x \in [0,1]^2} f(x),$$
$$|f(x) - f(y)| \leq \kappa \|x - y\|_2, \forall x, y \in [0,1]^2,$$
$$\text{piecewise affine on simplicial complex partition} \Big\}.$$

A simplicial complex partition divides $[0,1]^2$ into N equally sized triangles; see Section 5 for details and the fact that optimization over F can be reduced to solving a finite-dimensional convex problem. As discussed there, F can be viewed as an approximation, introduced for computational reasons, of the class obtained from F by relaxing the piecewise affine restriction. We also adopt the penalty term $\pi(f) = \lambda \sum_{k=1}^{N} \|g_i\|_1$, where g_i is the gradient of the ith affine function defining f. In the results reported here, $\kappa = 100$ with $\bar{x} = (0.4702, 0.4657)$ and $\bar{y} = (0.7746, 0.7773)$. We observe that F is misspecified as f^0 is not Lipschitz continuous.

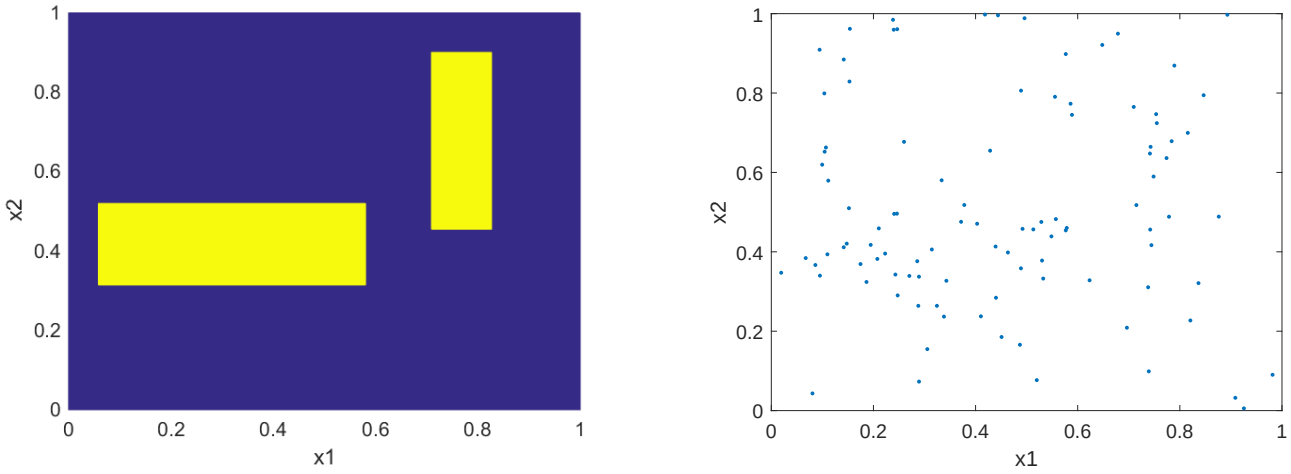

Figure 1: Top view of actual density (left) and sample of size $n = 100$ (right).

Figure 2 illustrates the effect of including the argmax-constraint for the case with $\lambda = 0.05$, $\alpha = 0.0001$, $\beta = 10000$, and $N = 200$. In the left portion of the figure, the argmax-constraint is not used and, visually, the errors are large. In the right portion, the argmax-constraint is included and indications of the actual density emerges. This and other experiments show that argmax-constraints regularize the estimates in some sense.

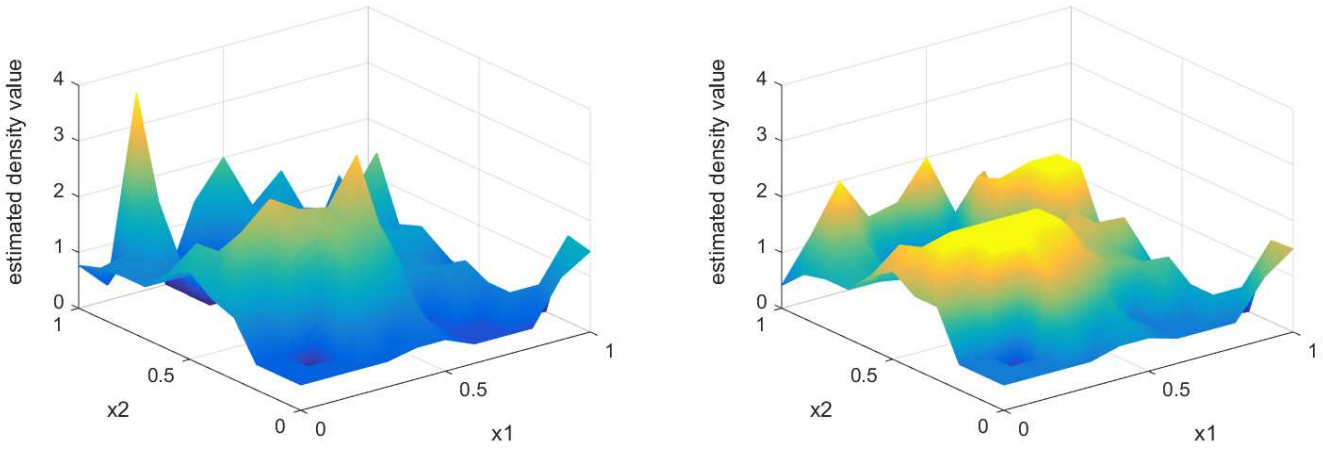

Figure 2: Estimates using $n = 100$ without (left) and with (right) argmax-constraint.

If α is increased to 0.3075 and β lowered to 4.5, i.e., 50% below and above the lowest and highest point of f^0, the estimate with argmax-constraint is slightly improved; see Figure 3(left) for a top-view of the resulting density. The estimates are quite insensitive to the choice of $\bar{x}$ and $\bar{y}$. Over 25 replications with $\bar{x}$ randomly selected from the box constituting the left portion of $\text{argmax}_{x\in[0,1]^2} f^0(x)$ and with $\bar{y}$ randomly selected from the right box, 22 estimates resemble strongly that in Figure 3(left). The remaining three blur together the two peaks of f^0. Still, the KL-divergence between $\hat{f}^n$ and f^0 remains close: the mean across the 25 replications is 0.177 and the standard deviation is 0.005. Naturally, a sample size of $n = 1000$ improves the estimates significantly; see Figure 3(right), where now $\lambda = 0.02$ and $N = 800$ are used.

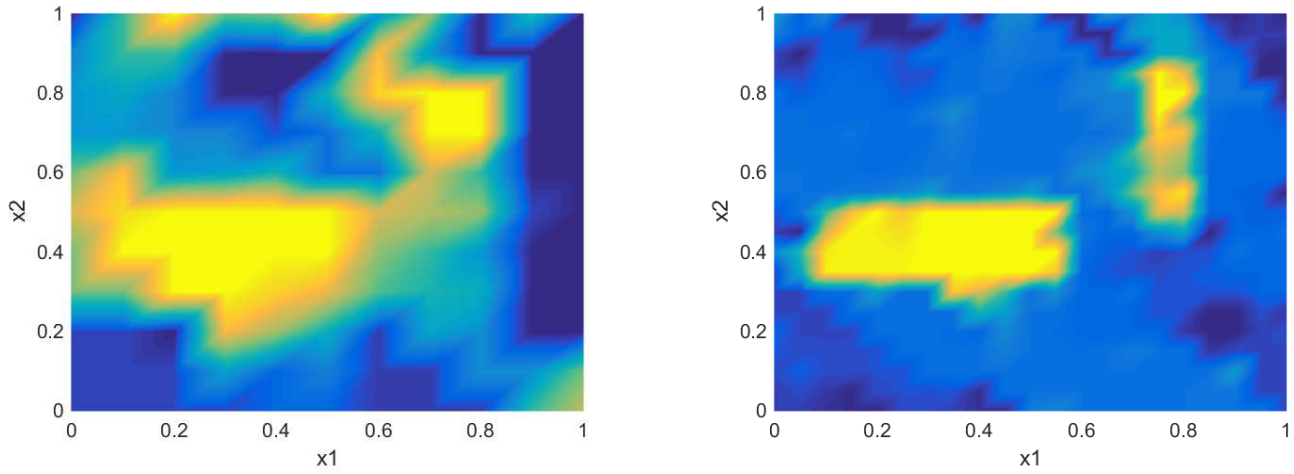

Figure 3: Estimates using sample size $n = 100$ (left) and $n = 1000$ (right).

partition	without penalty ($\lambda = 0$)			with penalty ($\lambda > 0$)		
size N	$n = 100$	$n = 1000$	$n = 10000$	$n = 100$	$n = 1000$	$n = 10000$
200	0.7	0.8	1.0	1.0	1.0	1.3
800	1.7	1.7	1.9	10.4	10.3	9.6
3200	6.6	11.7	14.8	38.5	29.1	22.0

Table 1: Computing times in seconds.

Table 1 summarizes typical computing times on a 2.60GHz laptop using IPOPT [48] under varying partition size N, sample size n, and penalty parameter; α and β are as before. The solver is not tuned for the specific problem instances and times can certainly be improved. In most cases, the run times are at most a few seconds. Interestingly, they are nearly constant in the sample size n as the size of the optimization problem is independent of n; see Section 5.2. Though, run times grow with partition size N. We observe that a piecewise affine density on a partition of $[0,1]^2$ with size N has $3N$ parameters that needs to be optimized. Thus, the last row in the table implies overfitting to some extent. The longer run times with penalties ($\lambda > 0$) are caused by additional optimization variables introduced in implementation of the nonsmooth penalty term. There are well-known techniques for mitigating this effect, but they are not explored here. Still, the table indicates the level of computational complexity for constrained M-estimator of this kind. Section 5 includes further discussion.

3 Existence and Consistency

After defining the aw-distance and establishing preliminary properties, this section turns to the main results on existence and consistency of estimators.

3.1 Attouch-Wets Distance

Throughout, we consider functions defined on a nonempty and closed set $S \subset \mathbb{R}^d$, which may be the whole of $\mathbb{R}^d$. In the density setting, S could be thought of as a support. However, we permit densities to have the value zero, so prior knowledge of the support is not required. The class F^n in (1) is viewed as a subset of the (extended real-valued) usc functions on S, which is denoted by

$$\text{usc-fcns}(S) = \{f : S \to \overline{\mathbb{R}} \mid f \text{ usc and } f \not\equiv -\infty\}, \text{ with } \overline{\mathbb{R}} = [-\infty, \infty].$$

Thus, $f \in \text{usc-fcns}(S)$ if and only if the *hypograph* $\text{hypo}\, f = \{(x,\alpha) \in S \times \mathbb{R} \mid f(x) \geq \alpha\}$ is a nonempty closed subset of $\mathbb{R}^d \times \mathbb{R}$. The class of usc functions is rich enough for most applications. We equip usc-fcns(S) with the aw-distance, which quantifies the distance between hypographs. Figure 4 shows $\text{hypo}\, f^n$ with shading and it appears "close" to $\text{hypo}\, f$. Specifically, let $\text{dist}(z, A)$

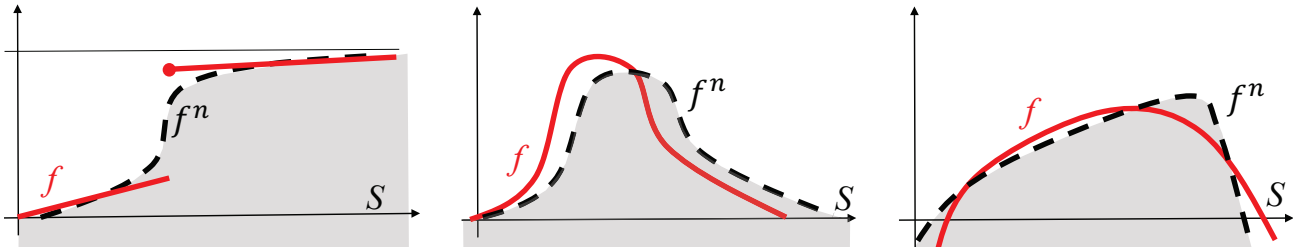

Figure 4: Hypographs of distribution (left), density (middle), and regression functions (right).

be the usual point-to-set distance between a point $z \in \mathbb{R}^d \times \mathbb{R}$ and a set $A \subset \mathbb{R}^d \times \mathbb{R}$; any norm $\|\cdot\|$ can be used. Let $z^{\mathrm{ctr}} \in S \times \mathbb{R}$. The choice of norm and z^{ctr} influence the numerical value of the aw-distance, but the resulting topology on usc-fcns(S) remains unchanged and thus all the stated results as well. For $f, g \in$ usc-fcns(S), the *aw-distance* is defined as

$$\mathbb{d}(f,g) = \int_0^\infty \mathbb{d}_\rho(f,g) e^{-\rho} d\rho,$$

where, for $\rho \geq 0$,

$$\mathbb{d}_\rho(f,g) = \max\Big\{ \big|\operatorname{dist}(z, \operatorname{hypo} f) - \operatorname{dist}(z, \operatorname{hypo} g)\big| \;\Big|\; \|z - z^{\mathrm{ctr}}\| \leq \rho \Big\}.$$

Indeed, (usc-fcns(S), $\mathbb{d}$) is a complete separable metric space, for which closed and bounded subsets are compact [29, Prop. 4.45, Thm. 7.58]. Boundedness can be verified by the inequality $\mathbb{d}(f,g) \leq 1 + \max\{\operatorname{dist}(z^{\mathrm{ctr}}, \operatorname{hypo} f), \operatorname{dist}(z^{\mathrm{ctr}}, \operatorname{hypo} g)\}$ [30, Prop. 3.1]. The aw-distance metrizes *hypo-convergence*: for $f^n, f \in$ usc-fcns(S),

$$\begin{aligned} f^n \text{ hypo-converges to } f &\iff \operatorname{hypo} f^n \text{ set-converges to } \operatorname{hypo} f \\ &\iff \begin{cases} \forall x^n \to x, \;\; \operatorname{limsup} f^n(x^n) \leq f(x) \\ \forall x \;\; \exists x^n \to x, \;\; \operatorname{liminf} f^n(x^n) \geq f(x) \end{cases} \\ &\iff \mathbb{d}(f^n, f) \to 0; \text{ simply denoted by } f^n \to f. \end{aligned} \tag{3}$$

Set-convergence is in the sense of Painlevé-Kuratowski[2]; see [29, Ch. 7].

Distribution functions hypo-converge if and only if they converge weakly [38, 37, 35] as illustrated in Figure 4(left). Figure 4(middle, right) hints to the fact that modes and maximizers of hypo-converging densities and regression functions converge to those of limiting functions; see Section 3.4.

[2]The *outer limit* of a sequence of sets $\{A^n, n \in \mathbb{N}\}$ in a topological space, denoted by $\operatorname{OutLim} A^n$, is the collection of points to which a subsequence of $\{a^n \in A^n, n \in \mathbb{N}\}$ converges. The *inner limit*, denoted by $\operatorname{InnLim} A^n$, is the collection of points to which a sequence $\{a^n \in A^n, n \in \mathbb{N}\}$ converges. If both limits exist and are equal to A, we say that $\{A^n, n \in \mathbb{N}\}$ *set-converges* to A and write $A^n \to A$ or $\operatorname{Lim} A^n = A$.

In general, $f^n \to f$ does not guarantee pointwise convergence; only limsup $f^n(x) \leq f(x)$ holds for all $x \in S$ by (3). This issue surfaces in the analysis of (semi)continuity properties of functions on usc-fcns(S). For $\bar{x} \in S$ and $\rho \geq 0$, let $\mathbb{B}(\bar{x}, \rho) = \{x \in S \mid \|\bar{x} - x\| \leq \rho\}$. We recall that $\{f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S) is *equi-usc* at $\bar{x} \in S$ when liminf $f^n(\bar{x}) \to \infty$ or when for every $\rho, \varepsilon \in (0, \infty)$, there exists $\bar{n} \in \mathbb{N}$ and $\delta > 0$ such that

$$\sup_{x \in B(\bar{x}, \delta)} f^n(x) \leq \max\{f^n(\bar{x}) + \varepsilon, -\rho\} \text{ for all } n \geq \bar{n}.$$

A class $F \subset$ usc-fcns(S) is equi-usc at $\bar{x} \in S$ when every sequence $\{f^n \in F, n \in \mathbb{N}\}$ is equi-usc at $\bar{x}$. The main consequence of this property is that hypo-convergence implies pointwise convergence [29, Thm. 7.10]:

3.1 Proposition (pointwise convergence). *If $\{f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S) is equi-usc at $\bar{x} \in S$, then $f^n \to f \in$ usc-fcns(S) implies $f^n(\bar{x}) \to f(\bar{x})$.*

Although the property is nontrivial, many interesting classes of functions are equi-usc at all, or "most," points in S as seen next. Let $\operatorname{int} A$ denote the interior of $A \subset \mathbb{R}^d$. A log-concave function $f = e^g$ for some concave function $g : S \to \overline{\mathbb{R}}$.

3.2 Proposition (sufficient conditions for equi-usc). *Any one of the following conditions suffice for the functions $\{f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S) to be equi-usc at $\bar{x} \in S$.*

(i) The functions are nonnegative, $f^n \to f \in$ usc-fcns(S), and $f(\bar{x}) = 0$.

(ii) The functions are concave, $f^n \to f \in$ usc-fcns(S), and $\bar{x} \in \operatorname{int}\{x \in S \mid f(x) > -\infty\}$.

(iii) The functions are log-concave, $f^n \to f \in$ usc-fcns(S), and $\bar{x} \in \operatorname{int}\{x \in S \mid f(x) > 0\}$.

(iv) The functions are nondecreasing[3] (alternatively, nonincreasing), $f^n \to f \in$ usc-fcns(S), and f is continuous at $\bar{x} \in \operatorname{int} S$.

(v) The functions are locally Lipschitz continuous at $\bar{x}$ with common modulus, i.e., there exist $\delta > 0$ and $\kappa \in [0, \infty)$ such that $|f^n(x) - f^n(\bar{x})| \leq \kappa \|x - \bar{x}\|$ for all $x \in \mathbb{B}(\bar{x}, \delta)$ and $n \in \mathbb{N}$.

Although the aw-distance cannot generally be related to some of the other common metrics, the Hellinger and L^2 distances tend to zero whenever the aw-distance vanishes under equi-usc and integrability assumptions.

3.3 Proposition (connections with other metrics). *Suppose that $\{f, f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S), $f^n \to f$, and for some (measurable) $g : S \to [0, \infty]$, $|f^n(x)| \leq g(x)$ for all $x \in S$ and $n \in \mathbb{N}$. Then,*

[3] Monotonicity of functions on S are always with respect to the partial order induced by inequalities interpreted componentwise, i.e., $f \in$ usc-fcns(S) is nondecreasing (nonincreasing) if $x \leq y$ implies $f^n(x) \leq (\geq) f^n(y)$.

(i) $L_P^2(f^n, f) = \int \big(f^n(x) - f(x)\big)^2 dP(x) \to 0$ *provided* $\int g^2(x)dP(x) < \infty$ *and* $\{f^n, n \in I\!N\}$ *is equi-usc at* P*-a.e.* $x \in S$*;*

(ii) $H^2(f^n, f) = \frac{1}{2}\int(\sqrt{f^n(x)} - \sqrt{f(x)})^2 dx \to 0$ *provided that* $f^n \geq 0$, $\int g(x)dx < \infty$, *and* $\{f^n, n \in I\!N\}$ *is equi-usc at Lebesgue-a.e.* $x \in S$.

3.2 Existence

Our first main result establishes that existence of an estimator reduces to having a semi-continuity property for the loss and penalty functions and a closed and bounded class of functions in the aw-distance.

We recall that a function $\varphi : F \to \overline{I\!R}$ defined on a closed subset F of usc-fcns(S) is lower-semicontinuous (lsc) if $\liminf \varphi(f^n) \geq \varphi(f)$ for all $f^n \in F \to f$. To clarify earlier notation[4], let $\varepsilon\text{-}\operatorname{argmin}_{f\in F} \varphi(f) = \{f \in F \mid \varphi(f) \leq \inf_{g\in F} \varphi(g) + \varepsilon\}$.

Although our focus is on the existence of M-estimators, i.e., minimizers of losses under an empirical distribution, occasionally we consider general distributions and thereby also treat approximation problems. We consider the following general setting [29, Ch. 14]: For a closed $F \subset$ usc-fcns(S) and a complete probability space $(S^0, \mathcal{B}^0, P^0)$, with $S^0 \subset I\!R^{d_0}$, we say that $\psi : S^0 \times F \to \overline{I\!R}$ is a *random lsc function* if for all $x \in S^0$, $\psi(x, \cdot)$ is lsc and ψ is measurable with respect to the product sigma-algebra[5] on $S^0 \times F$. A random lsc function $\psi : S^0 \times F \to \overline{I\!R}$ is *locally inf-integrable* if for all $f \in F$ there exists $\rho > 0$ such that[6] $\int \inf_{g\in F}\{\psi(x, g) \mid \mathit{d}\!\!\!^{-}(f, g) \leq \rho\} dP^0(x) > -\infty$.

3.4 Theorem (existence of approximation). *Suppose that* $\varepsilon \geq 0$ *and* F *is a nonempty, closed, and bounded subset of usc-fcns*(S) *and* $(S^0, \mathcal{B}^0, P^0)$ *is a complete probability space. If* $\psi : S^0 \times F \to \overline{I\!R}$ *is a locally inf-integrable random lsc function and* $\pi : F \to (-\infty, \infty]$ *is lsc, then*

$$\varepsilon\text{-}\operatorname{argmin}_{f\in F} \int \psi(x, f)dP^0(x) + \pi(f) \neq \emptyset$$

$$\text{and } \inf_{f\in F} \int \psi(x, f)dP^0(x) + \pi(f) > -\infty.$$

3.5 Corollary (existence of estimator). *Suppose that* $\varepsilon \geq 0$, $\{x^j \in I\!R^{d_0},\ j = 1, \ldots, n\}$, *and* F *is a nonempty, closed, and bounded subset of usc-fcns*(S). *If* $\pi : F \to (-\infty, \infty]$ *and* $\psi(x^j, \cdot) : F \to (-\infty, \infty]$ *are lsc for all* j, *then*

$$\varepsilon\text{-}\operatorname{argmin}_{f\in F} \frac{1}{n}\sum_{j=1}^{n} \psi(x^j, f) + \pi(f) \neq \emptyset \text{ and } \inf_{f\in F} \frac{1}{n}\sum_{j=1}^{n} \psi(x^j, f) + \pi(f) > -\infty.$$

[4]Throughout we use the common extended real-valued calculus: $0 \cdot \infty = 0$, $\alpha \cdot \infty = \infty$ for $\alpha > 0$, $\alpha + \infty = \infty$ for $\alpha \in \overline{I\!R}$, and $\alpha - \infty = -\infty$ for $\alpha \in [-\infty, \infty)$; see [29, Sec. 1.E].

[5]For F, we adopt the Borel sigma-algebra under $\mathit{d}\!\!\!^{-}$.

[6]With $\infty - \infty = \infty$, the integral of any measurable function is well-defined. In particular, the present integrand is measurable [29, Thm. 14.37], [12, Prop. 6.3].

For F to be bounded it suffices that there are $x \in S$ and $\alpha \in \mathbb{R}$ such that for all $f \in F$, $f(x) \geq \alpha$, which becomes trivial for densities and distribution functions. As we see below, the condition can sometimes be removed. Many natural classes are closed as indicated in the introduction and detailed in Section 4. The common penalty function $\pi(f) = \sup_{x \in S} |f(x)|$ is lsc (cf. Proposition 4.7). Familiar loss functions satisfy the lsc requirement too, at least under certain assumptions. Several examples are furnished including some involving support vector machines (SVM). The class in the next corollary considers concave classifiers in a "band" that are also subject to constraints on the location of level-sets.

3.6 Corollary (existence of concave SVM classifier). *For $g : S \to (-\infty, \infty]$, $h \in$ usc-fcns(S), $\alpha \in \mathbb{R}$, and $C \subset \mathbb{R}^d$, suppose that $\{y^j \in \{-1, 1\}, x^j \in \operatorname{int} S,\ j = 1, \ldots, n\}$ and $F = \{f \in$ usc-fcns$(S) \mid f$ concave, $g(x) \leq f(x) \leq h(x)\ \forall x \in S,\ C \subset \operatorname{lev}_{\geq \alpha} f\}$. Then, as long as F is nonempty,*

$$\operatorname{argmin}_{f \in F} \frac{1}{n} \sum_{j=1}^{n} \max\left\{0, 1 - y^j f(x^j)\right\} \neq \emptyset.$$

When classification errors of different types need to be treated separately, a Neyman-Pearson model leads to the following setting [4, 5, 40].

3.7 Corollary (existence of robust Neyman-Pearson classifier). *Suppose that $\{x^j \in S,\ j = 1, \ldots, n\}$ and $\{z^i \in S,\ i = 1, \ldots, m\}$ are associated with $+1$ and -1 labels, respectively, and F is a nonempty closed subset of* usc-fcns(S). *Then, for open sets $\{Z^i \subset \mathbb{R}^d, i = 1, \ldots, m\}$, with $z^i \in Z^i$,*

$$\operatorname{argmin}_{f \in F} \left\{ \frac{1}{n} \sum_{j=1}^{n} \max\left\{0, 1 - f(x^j)\right\} \;\middle|\; f(z) \leq 0\ \forall z \in Z^i, i = 1, \ldots, m \right\} \neq \emptyset.$$

The corollary establishes the existence of an estimator, defined by a broad class F, that minimizes hinge loss across the $+1$ labels and tolerates no training error across the -1 labels even after perturbations within sets Z^i.

3.8 Corollary (existence of ML estimator). *If F is a nonempty closed subset of* usc-fcns(S) *consisting of nonnegative functions, $\varepsilon \geq 0$, $\{x^j \in S,\ j = 1, \ldots, n\}$, and $f(x^j) < \infty$ for all j and $f \in F$, then*

$$\varepsilon\text{-}\operatorname{argmin}_{f \in F} -\frac{1}{n} \sum_{j=1}^{n} \log f(x^j) \neq \emptyset \text{ and } \inf_{f \in F} -\frac{1}{n} \sum_{j=1}^{n} \log f(x^j) > -\infty.$$

We observe that the corollary actually applies to any $f : S \to [0, \infty]$ and not only densities[7]. This fact is beneficial in analysis of estimators for which the integral-to-one constraint is relaxed,

[7] We extend $\alpha \mapsto \log \alpha$ to $[0, \infty]$ by assigning the end points $-\infty$ and ∞, respectively.

for example, due to computational concerns. Nevertheless, the constraint enters in many settings and needs a closer examination.

If F is the class of normal densities with mean zero and positive standard deviation, then F is not closed because there is a sequence in F hypo-converging to a degenerate density with zero standard deviation. Similarly, if F is the class of normal densities with standard deviation one, then closedness fails again since one can construct densities in F hypo-converging to the zero function. Also classes of bounded densities on a compact set S may not be closed. Elimination of such pathological cases is required for a class of densities to be closed. Proposition 2.1 furnishes a concrete example, while Proposition 4.8 shows that if F is equi-usc at Lebesgue-a.e. $x \in S$ and an integrability condition holds, then $\int f(x)dx = 1$ is closed under hypo-convergence. The log-concave class exhibits an equi-usc property as established in Proposition 3.2. It is therefore not surprising that the ML estimator over this class exists under a mild condition on the sample [14]; see Proposition 6.2 below.

3.9 Corollary (LS regression). *Suppose that F is a nonempty closed subset of usc-fcns(S). If $\{y^j \in I\!R, x^j \in S, j = 1, \ldots, n\}$ and F is equi-usc at x^j, $j = 1, \ldots, n$, then*

$$\operatorname{argmin}_{f \in F} \frac{1}{n} \sum_{j=1}^{n} (y^j - f(x^j))^2 \neq \emptyset.$$

Proposition 3.2 gives various sufficient conditions for a class of functions to be equi-usc. The concave functions are equi-usc at "most" points according to that proposition and a variant of LS regression that also includes pointwise upper and lower bound, for example introduced to engineer desirable estimates in high-dimensional settings, does indeed exists. This can be established using the same arguments as those supporting Corollary 3.6.

In special cases with relatively simple constraints such as only monotonicity or only convexity, existence of LS estimators are well-known; see [39, 41]. The key feature of these special cases is that they reduce in some sense to finite-dimensional problems expressed in terms of the heights $\theta_j = f(x^j), j = 1, \ldots, n$, and the limit of sequence of such heights generated by feasible functions can easily be shown to be extendable to a feasible function. In the presence of nontrivial constraints that impose restrictions on f at points other than the design points, the situation is more complicated and our systematic approach has merit. In particular, starting from a closed equi-usc class, one can build up closed equi-usc classes through set operations that preserve closedness such as intersections and thereby construct novel estimators that will exist by Corollary 3.9.

3.3 Consistency

Our second main result establishes that consistency follows essentially from lower-semicontinuity and one-sided integrability of the loss function and the closedness of the class under consideration.

3.10 Theorem (consistency). *Suppose that $X^1, X^2, \ldots$ are iid random vectors with values in $S^0 \subset \mathbb{R}^{d_0}$, F is a closed subset of usc-fcns(S), $\psi : S^0 \times F \to \overline{\mathbb{R}}$ is a locally inf-integrable random lsc function, and $\pi^n : F \to [0, \infty)$ satisfies $\pi^n(f^n) \to 0$ for every convergent sequence $\{f^n \in F, n \in \mathbb{N}\}$. Then, the following hold almost surely:*

(i) For all $\{\varepsilon^n \geq 0, n \in \mathbb{N}\} \to 0$,

$$\operatorname{OutLim}\Big(\varepsilon^n\text{-}\operatorname{argmin}_{f\in F} \frac{1}{n}\sum_{j=1}^{n} \psi(X^j, f) + \pi^n(f)\Big) \subset \operatorname{argmin}_{f\in F} \mathbb{E}[\psi(X^1, f)].$$

(ii) There exists $\{\varepsilon^n \geq 0, n \in \mathbb{N}\} \to 0$, such that

$$\Big(\varepsilon^n\text{-}\operatorname{argmin}_{f\in F} \frac{1}{n}\sum_{j=1}^{n} \psi(X^j, f) + \pi^n(f)\Big) \to \operatorname{argmin}_{f\in F} \mathbb{E}[\psi(X^1, f)]$$

provided that $\mathbb{E}[\psi(X^1, f)] < \infty$ for at least one $f \in F$ and F is bounded.

The first conclusion of Theorem 3.10 guarantees that every cluster point of sequences constructed from near-minimizers of $n^{-1}\sum_{j=1}^{n} \psi(X^j, \cdot) + \pi^n$ is contained in $\operatorname{argmin}_{f\in F} \mathbb{E}[\psi(X^1, f)]$ provided that ε^n vanishes.

Since $\operatorname{argmin}_{f\in F} \mathbb{E}[\psi(X^1, f)]$ may not be a singleton, especially under model misspecification, there might be a strict inclusion in the first conclusion. For example, let $S = S^0 = [0, 1]$, $F = \{f \mid f(x) = 1 \text{ for } x \in [0, 1), f(1) \in [1, 2]\}$, the actual density f^0 be uniform on S, and $\pi^n(f) = n^{-1} \sup_{x\in S} f(x)$. Then, almost surely, $\operatorname{argmin}_{f\in F} -n^{-1}\sum_{j=1}^{n} \log f(X^j) + \pi^n(f) = \{f^0\}$, a strict subset of $\operatorname{argmin}_{f\in F} \mathbb{E}[-\log f(X^1)] = F$. In this example, the difficulty is caused by effects on a set of Lebesgue measure zero. However, in more complicated situations, the concern may be more prevalent. An example is furnished by the same f^0, S, and S^0, but with $F = \{g^1, g^2\}$, where $g^1(x) = 1 + \delta$ for $x \in [0, 1/2]$ and $g^1(x) = 1 - \delta$ for $x \in (1/2, 1]$, and $g^2(x) = 1 - \delta$ for $x \in [0, 1/2]$ and $g^2(x) = 1 + \delta$ for $x \in (1/2, 1]$, where $\delta \in (0, 1)$, and $\pi^n(f) = n^{-1/2} f(0)$. The actual density f^0 is outside F. Then, almost surely, $\operatorname{OutLim}\{\operatorname{argmin}_{f\in F} -n^{-1}\sum_{j=1}^{n} \log f(X^j) + \pi^n(f)\} = \{g^2\}$, a strict subset of $\operatorname{argmin}_{f\in F} \mathbb{E}[-\log f(X^1)] = F$.

The second conclusion in Theorem 3.10 guarantees that if ε^n tends to zero sufficiently slowly, then the inclusion cannot be strict; near-minimizers of $n^{-1}\sum_{j=1}^{n} \psi(X^j, \cdot) + \pi^n$ set-converge to $\operatorname{argmin}_{f\in F} \mathbb{E}[\psi(X^1, f)]$. Thus, in this sense, estimators can converge to *any* function in the latter argmin.

A comparison with the common approach to consistency laid out, for example, in [44, Sec. 3.2.1] is illuminating. In our notation, [44, Cor. 3.2.3] states roughly that if (i) $n^{-1}\sum_{j=1}^{n} \psi(X^j, f)$ converges in probability to $\mathbb{E}[\psi(X^1, f)]$ uniformly in f across F, which is permitted to be *any* metric space, and (ii) $\mathbb{E}[\psi(X^1, \cdot)]$ has a well-separated (unique) minimizer f^0 on F, then $\hat{f}^n$

converges in probability to f^0. The ability to handle an arbitrary metric space is an advantage over Theorem 3.10, but also burdens the user with verifying the well-separability of f^0 in the chosen metric. We do not insist on a unique minimizer as discussed above. The required uniform weak law of large numbers would typically need $\psi(X^1, f)$ to be integrable. In contrast, Theorem 3.10 insists only on a one-sided integrability condition, which is trivially satisfied when $\psi(x, f)$ is uniformly bounded from below across $x \in S^0$ and $f \in F$ as would be the case for hinge-loss, least-squares, and other common loss functions.

3.11 Corollary (consistency for concave SVM classifier). *For $g : \mathbb{R}^d \to (-\infty, \infty]$, $h \in$ usc-fcns($\mathbb{R}^d$), $\gamma \in \mathbb{R}$, and $C \subset \mathbb{R}^d$, suppose that $(X^1, Y^1), (X^2, Y^2), \ldots$ are iid random vectors in $\mathbb{R}^d \times \{-1, 1\}$ and $F = \{f \in$ usc-fcns($\mathbb{R}^d$) $\mid f$ concave, $g(x) \le f(x) \le h(x)\ \forall x \in \mathbb{R}^d$, $C \subset \mathrm{lev}_{\ge\gamma} f\}$.*

If $\{\varepsilon^n \ge 0, n \in \mathbb{N}\} \to 0$ and

$$\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f\in F} \frac{1}{n}\sum_{j=1}^{n} \max\{0, 1 - Y^j f(X^j)\},$$

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

$$f^\star \in \operatorname{argmin}_{f\in F} \mathbb{E}\big[\max\{0, 1 - Y^1 f(X^1)\}\big].$$

Moreover, for a subsequence $\{n_k, k \in \mathbb{N}\}$ with $\hat{f}^{n_k} \to f^\star$ and $\beta < \alpha \in \mathbb{R}$,

$$\mathrm{OutLim}_k \left(\mathrm{lev}_{\ge\alpha} \hat{f}^{n_k}\right) \subset \mathrm{lev}_{\ge\alpha} f^\star \text{ and } \mathrm{InnLim}_k \left(\mathrm{lev}_{\ge\beta} \hat{f}^{n_k}\right) \supset \mathrm{lev}_{\ge\alpha} f^\star.$$

We note that the upper level-sets of $\hat{f}^n$, which are central in the practical use of the classifier (especially for $\alpha = 0$), indeed approximate the "true" level-set $\mathrm{lev}_{\ge\alpha} f^\star$. Without additional assumptions, we are unable to permit $\beta = \alpha$ because it is fundamentally difficult to estimate $\mathrm{lev}_{\ge\alpha} f^\star$ when $f^\star(x) = \alpha$ on a set of positive measure. For consistency of SVM defined over a subset of a reproducing kernel Hilbert space, we refer to [43].

The Kullback-Leibler divergence

$$K(g; f) = \int g(x)\big[\log g(x) - \log f(x)\big]dx \text{ for (measurable) } f, g : S \to [0, \infty]$$

enters in ML estimation of densities.

3.12 Corollary (consistency in ML estimation). *Suppose that $X^1, X^2, \ldots$ are iid random vectors, each distributed according to a density $f^0 : S \to [0, \infty]$, F is a closed subset of usc-fcns(S) with nonnegative functions, and for every $f \in F$ there exists $\rho > 0$ such that $\mathbb{E}[\sup_{g\in F}\{\log g(X^1) \mid d\!\!l(f, g) \le \rho\}] < \infty$. If $\{\varepsilon^n \ge 0, n \in \mathbb{N}\} \to 0$ and*

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

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

$$f^\star \in \operatorname{argmin}_{f \in F} K(f^0; f).$$

Under the additional assumption that F *contains only densities and* $f^0 \in F$*, we also have that* $f^\star(x) = f^0(x)$ *for Lebesgue-a.e.* $x \in S$*.*

It is obvious that when there exists an $\alpha \in \mathbb{R}$ such that $f(x) \le \alpha$ for all $f \in F$, then the expectation assumption is satisfied. In particular, such an α exists if for some $\kappa \in [0, \infty)$ the class $F \subset \{f : S \to [0, \infty] \mid \int f(x)dx = 1,\ |f(x) - f(y)| \le \kappa \|x - y\|_2\ \forall x, y \in S\}$. Alternatively, if X^1 is integrable and there exist $\alpha, \beta \in \mathbb{R}$ such that $f(x) \le \exp(\alpha + \beta\|x\|_\infty)$ for all $f \in F$, then again the expectation assumption in the corollary is satisfied.

We next turn the attention to LS regression. Suppose that we are given the random design model

$$Y^j = f^0(X^j) + Z^j, \quad j = 1, 2, \ldots,$$

where the iid random vectors $X^1, X^2, \ldots$ take values in the closed set $S \subset \mathbb{R}^d$, the iid zero-mean and finite-variance random variables $Z^1, Z^2, \ldots$ are also independent of $X^1, X^2, \ldots$, and $f^0 : S \to \mathbb{R}$ is an unknown function to be estimated based on observations of (X^1, Y^1). Let

$$L_P^2(f, g) = \int \big(f(x) - g(x)\big)^2 dP(x),$$

where P is the distribution of X^1. Consistency in the aw-distance is stated next; see [18] for consistency in the empirical L^2 sense.

3.13 Corollary (consistency in LS regression). *Suppose that* $\{\varepsilon^n \ge 0, n \in \mathbb{N}\} \to 0$ *and* F *is a closed subset of usc-fcns*(S) *equi-usc at every* $x \in S$*. For the random design model above and*

$$\hat{f}^n \in \varepsilon^n\text{-}\operatorname{argmin}_{f \in F} \frac{1}{n} \sum_{j=1}^{n} \big(Y^j - f(X^j)\big)^2,$$

we have, almost surely, that every cluster point $f^\star$ *of* $\{\hat{f}^n, n \in \mathbb{N}\}$ *satisfies*

$$f^\star \in \operatorname{argmin}_{f \in F} L_P^2(f, f^0).$$

If $\inf_{f \in F} \mathbb{E}[(Y^1 - f(X^1))^2] < \infty$*, which occurs in particular when* $f^0 \in F$*, then* $\{\hat{f}^n, n \in \mathbb{N}\}$ *has at least one cluster point.*

When $f^0 \in F$*, we also have that* $f^\star(x) = f^0(x)$ *for* P*-a.e.* $x \in S$*.*

We next turn to consistency in the presence of sieves, i.e., the class of functions F^n varies with n. The importance of sieves is well-documented and prior studies include [11, 10, 20, 19, 9, 6]; see also [47, Thms. 8.4 and 8.12].

3.14 Theorem (consistency; sieves). *Suppose that $X^1, X^2, \ldots$ are iid random vectors with values in $S^0 \subset \mathbb{R}^{d_0}$, F is a closed subset of usc-fcns(S), $F^n \subset F$, $\psi : S^0 \times F \to \overline{\mathbb{R}}$ is a locally inf-integrable random lsc function, $\pi^n : F \to [0, \infty)$ satisfies $\pi^n(f^n) \to 0$ for every convergent sequence $\{f^n \in F, n \in \mathbb{N}\}$, and $\delta > 0$. If $\{\varepsilon^n \geq 0, n \in \mathbb{N}\} \to 0$, then*

$$\begin{aligned}\operatorname{OutLim} \Big(\varepsilon^n\text{-}\operatorname{argmin}_{f \in F_\delta^n} \frac{1}{n} \sum_{j=1}^{n} \psi(X^j, f) + \pi^n(f) \Big) \\ \subset \big\{ f \in F_\delta^\infty \;\big|\; \mathbb{E}[\psi(X^1, f)] \leq \inf_{g \in \operatorname{Lim} F^n} \mathbb{E}[\psi(X^1, g)] \big\} \text{ a.s.},\end{aligned}$$

where $F_\delta^n = \{f \in F \mid \inf_{g \in F^n} d(f, g) \leq \delta\}$ and F_δ^∞ is defined similarly with F^n replaced by $\operatorname{Lim} F^n$. In particular, if $\operatorname{Lim} F^n = F$, then the right-hand side of the inclusion equals $\operatorname{argmin}_{f \in F} \mathbb{E}[\psi(X^1, f)]$.

The assumptions of the theorem are nearly identical to those of Theorem 3.10. The main difference is that consistency is ensured for estimators that are near-minimizers of a slightly *relaxed* problem over the class F_δ^n and not over F^n. This relaxation is potentially beneficial from a computationally point of view (see Section 5.1).

Theorem 3.14 guarantees that estimators selected from such relaxed classes will be consistent in some sense. Specifically, every cluster point of the estimators is at least as "good" as $\inf_{g \in \operatorname{Lim} F^n} \mathbb{E}[\psi(X^1, g)]$ and is also in F_δ^∞. If F^n eventually "fills" F, consistency takes place in the usual sense.

To illustrate one application area, we specialize the theorem for ML estimation of densities, while retaining some of its notation.

3.15 Corollary (consistency in ML estimation; sieves). *Suppose that $X^1, X^2, \ldots$ are iid random vectors, each distributed according to a density $f^0 : S \to [0, \infty]$, F is a closed subset of usc-fcns(S) consisting of densities, $F^n \subset F$, and for every $f \in F$ there exists $\rho > 0$ such that $\mathbb{E}[\sup_{g \in F}\{\log g(X^1) \mid d(f, g) \leq \rho\}] < \infty$. If $\delta > 0$, $\{\varepsilon^n \geq 0, n \in \mathbb{N}\} \to 0$, $f^0 \in \operatorname{Lim} F^n$, and*

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

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

$$K(f^0; f^\star) = 0 \text{ and } f^\star \in F_\delta^\infty.$$

Thus, $f^\star(x) = f^0(x)$ for Lebesgue-a.e. $x \in S$.

3.4 Plug-In Estimators

Among the many plug-in estimators that can be constructed from density estimators, 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. Similarly, plug-in estimators of "peaks" of regression functions and level-sets of classifiers will also be consistent. Maxima and maximizers of regression functions are important, especially in engineering design where "surrogate models" are built using regression and that are subsequently maximized to find an optimal design or decision.

We recall that $\varepsilon\text{-}\operatorname{argmax}_{x\in S} f(x) = \{y \in S \mid f(y) \geq \sup_{x\in S} f(x) - \varepsilon\}$ for $\varepsilon \geq 0$ and $f : S \to \overline{\mathbb{R}}$. Thus, $f(x^\star) = \infty$ when $x^\star \in \varepsilon\text{-}\operatorname{argmax}_{x\in S} f(x)$ and $\sup_{x\in S} f(x) = \infty$. If f is a density, then $\operatorname{argmax}_{x\in S} f(x)$ is the set of *modes* of f, $\delta\text{-}\operatorname{argmax}_{x\in S} f(x)$ is a set of *near-modes*, and $\operatorname{lev}_{\geq\alpha} f$ is a set of *high-likelihood events*. We stress that modes are defined here as *global* maximizers of densities. Extension to a more inclusive definition is possible but omitted.

3.16 Theorem (plug-in estimators of modes and related quantities). *Suppose that estimators $\hat{f}^n \to f^0$ almost surely, with estimates being functions in usc-fcns(S). If $\{\delta^n \geq 0,\ n \in \mathbb{N}\} \to \delta$ and $\{\alpha^n \in \overline{\mathbb{R}},\ n \in \mathbb{N}\} \to \alpha$, then the plug-in estimators*

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

are consistent in the sense that almost surely $\delta\text{-}\operatorname{argmax}_{x\in S} f^0(x)$ and $\operatorname{lev}_{\geq\alpha} f^0$ contain every cluster point of $\{\hat{m}^n, n \in \mathbb{N}\}$ and $\{\hat{l}^n, n \in \mathbb{N}\}$, respectively.

Moreover, if there is a compact $B \subset S$ such that for all n $\operatorname{argmax}_{x\in S} \hat{f}^n(x) \cap B \neq \emptyset$ almost surely, then the plug-in estimator

$$\hat{h}^n = \sup_{x\in S} \hat{f}^n(x) \to \sup_{x\in S} f^0(x) \text{ almost surely.}$$

The theorem provides foundations for a rich class of *constrained* estimators for modes, near-modes, height of modes, and high-likelihood events and similar quantities for regression functions and classifiers. We observe that the theorem holds even if f^0 fails to have a unique maximizer. 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 and related quantities without additional assumptions.

4 Closed Classes

The central technical challenge associated with applying our existence and consistency theorems is often to establish that the class of functions under consideration is a closed subset of usc-fcns(S). The analysis is significantly simplified by the fact that any intersection of closed sets is also closed. Thus, it suffices to examine each *individual* requirement of a class separately.

It is well known that the limit of a hypo-converging sequence of concave functions must also be concave and thus the class of concave functions is closed [29, Prop. 4.15]. In this section, we provide numerous results for other classes. We note that S is necessarily convex when $f \in$ usc-fcns(S) is convex, concave, or log-concave.

4.1 Proposition (convexity and (log-)concavity). *For $\{f, f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S) and $f^n \to f$, we have:*

(i) If $\{f^n, n \in \mathbb{N}\}$ are concave, then f is concave. Moreover, if the functions are finite-valued, $\kappa \geq 0$, and $\|v\|_2 \leq \kappa$ for every subgradient $v \in \partial f^n(x)$ and $x \in S$, then $\|v\|_2 \leq \kappa$ for every $v \in \partial f(x)$ and $x \in S$.

(ii) If $\{f^n \geq 0, n \in \mathbb{N}\}$ are log-concave, then f is log-concave.

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

The additional assumption about int S being nonempty for the convex case is caused by the fact that the aw-distance is inherently tied to hypographs, which makes the treatment of convex functions slightly more delicate than that of concave functions.

Transformations of convex and concave functions beyond the log-concave case lead to the rich class of s-concave densities; see for example [42, 22].

4.2 Proposition (monotone transformations). *For a continuous nondecreasing function $h_0 : \mathbb{R} \to \overline{\mathbb{R}}$, let $h : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ have $h(y) = h_0(y)$ if $y \in \mathbb{R}$, $h(-\infty) = \inf_{\bar{y} \in R} h_0(\bar{y})$, and $h(\infty) = \sup_{\bar{y} \in R} h_0(\bar{y})$. Then, for $\{g^n : S \to \overline{\mathbb{R}}, n \in \mathbb{N}\}$, with $h \circ g^n \in$ usc-fcns$(S) \to f \in$ usc-fcns(S), the following hold:*

(i) If $\{g^n,\ n \in \mathbb{N}\}$ are concave, then $f = h \circ g$ for some concave $g : S \to \overline{\mathbb{R}}$.

(ii) If $\{g^n,\ n \in \mathbb{N}\}$ are convex and int S *is nonempty, then $f = h \circ g$ for some convex $g : S \to \overline{\mathbb{R}}$.*

Since $h \circ g$ with h nonincreasing and g convex can be written as $\tilde{h} \circ \tilde{g}$ with $\tilde{h}$ nondecreasing and $\tilde{g}$ concave, the proposition also addresses nonincreasing functions and in fact all *s-concave* functions. This ensures closedness for classes of functions under such shape restrictions.

4.3 Proposition (monotonicity). *For $\{f, f^n, n \in \mathbb{N}\} \subset$ usc-fcns(S) and $f^n \to f$, we have:*

(i) If f^n is nondecreasing in the sense that $f^n(x) \leq f^n(y)$ for $x \in S$, $y \in$ int S*, with $x \leq y$, then f is also nondecreasing in the same sense.*

If S is a box[8]*, then* int S *can be replaced by S.*

(ii) If f^n is nonincreasing in the sense that $f^n(x) \geq f^n(y)$ for $x \in$ int S*, $y \in S$, with $x \leq y$, then f is also nonincreasing in the same sense.*

If S is a box, then int S *can be replaced by S.*

[8] A box in $\mathbb{R}^d$ is of the form $S = [\alpha_1, \beta_2] \times \ldots [\alpha_d, \beta_d]$, with $-\infty \leq \alpha_i < \beta_i \leq \infty$, where in the case of $\alpha_i = -\infty$ and $\beta_i = \infty$ the closed intervals are replaced by (half)open intervals. Its dimension is therefore d.

The limit of a hypo-converging sequence of nondecreasing functions is not necessarily nondecreasing for arbitrary S. Consider $S = \{(x_1, x_2) \in I\!R^2 \mid x_1 = x_2, 0 \leq x_1, x_2 \leq 1\} \cup \{(2,0)\}$, $f(x) = f^n(x) = 0$ if $x = (2,0)$, and $f(x) = 1$ and $f^n(x) = \min\{1, n(x_1 + x_2)\}$ otherwise. Clearly, $x = (0,0) \leq y = (2,0)$, 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, $f^n \to f$.

We recall that $f : S \to \overline{I\!R}$ is Lipschitz continuous with modulus κ when $|f(x) - f(y)| \leq \kappa\|x - y\|$ for all $x, y \in S$.

4.4 Proposition (Lipschitz continuity). *Suppose that $\{f, f^n, n \in I\!N\} \subset$ usc-fcns(S), $f^n \to f$, and $\{f^n, n \in I\!N\}$ are Lipschitz continuous with common modulus κ. Then, f is also Lipschitz continuous with modulus κ.*

4.5 Proposition (pointwise bounds). *Suppose that $g : S \to \overline{I\!R}$, $h \in$ usc-fcns (S), $\{f, f^n, n \in I\!N\} \subset$ usc-fcns(S), and $f^n \to f$. If $g(x) \leq f^n(x) \leq h(x)$ for all $n \in I\!N$ and $x \in S$, then $g(x) \leq f(x) \leq h(x)$ for all $x \in S$.*

A function $f : S \to \overline{I\!R}$ is in the class of multivariate totally positive functions of order two when $f(x)f(y) \leq f(\min\{x,y\})f(\max\{x,y\})$ for all $x, y \in S$; see for example [17]. The min and max are taken componentwise.

4.6 Proposition (multivariate total positivity of order two). *If $\{f^n, n \in I\!N\} \subset$ usc-fcns(S) is equi-usc at $\bar{x} \in S$, the functions f^n are multivariate totally positive of order two, and $f^n \to f \in$ usc-fcns(S), then f is multivariate totally positive of order two.*

Penalty terms and constraints are often defined in terms of sup-functions and integrals. Their (semi)-continuity properties are recorded next.

4.7 Proposition (lsc of sup-norm). *If $F \subset$ usc-fcns(S) and $g : \overline{I\!R} \to \overline{I\!R}$ is lsc[9], then $\pi : F \to \overline{I\!R}$ defined by $\pi(f) = \sup_{x \in S} g(f(x))$ is lsc.*

In particular, $f \mapsto \sup_{x \in S} |f(x)|$ is lsc because this corresponds to having $g(y) = |y|$ for $y \in I\!R$ and $g(y) = \infty$ for $y = -\infty$ and ∞ in the proposition.

4.8 Proposition (integral quantities). *If $\{f^n, n \in I\!N\} \subset$ usc-fcns(S) is equi-usc at Lebesgue-a.e. $x \in S$, $f^n \to f \in$ usc-fcns(S), and for some (measurable) $g : S \to [0, \infty]$, $|f^n(x)| \leq g(x)$ for all $x \in S$ and $n \in I\!N$, then*

(i) $\int f^n(x)dx \to \int f(x)dx$ provided $\int g(x)dx < \infty$;

(ii) $\int xf^n(x)dx \to \int xf(x)dx$ provided $\int \|x\|g(x)dx < \infty$.

We end the section with an example of approximating and/or evolving moment information in the definition of a function class.

[9] $g : \overline{I\!R} \to \overline{I\!R}$ is lsc if $\liminf g(y^n) \geq g(y)$ for every $y^n \to y \in \overline{I\!R}$.

4.9 Proposition (moment information.) *Suppose that $C \subset C^n \subset \mathbb{R}^d$ are closed, $F^0 \subset$ usc-fcns(S) is closed and equi-usc at every $x \in S$, and there is a function $g : S \to [0, \infty]$ with $\int \|x\| g(x) dx < \infty$ and $|f(x)| \leq g(x)$ for all $x \in S$ and $f \in F^0$. Let*

$$F = \Big\{ f \in F^0 \;\Big|\; \int x f(x) dx \in C \Big\} \text{ and } F^n = \Big\{ f \in F^0 \;\Big|\; \int x f(x) dx \in C^n \Big\}.$$

If C^n set-converges to C, then F^n set-converges to F.

5 Estimation Algorithm

For given data $x^1, \ldots, x^n \in S^0 \subset \mathbb{R}^{d_0}$, there are no general algorithms available for finding a function in

$$\varepsilon\text{-}\operatorname{argmin}_{f \in F} \frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f) + \pi(f). \tag{4}$$

In this section, we provide an algorithm for this purpose that combines the need for approximation of functions in usc-fcns(S) with the use of state-of-the-art solvers for finite-dimensional optimization.

Suppose that π^ν is an approximation of π and F^ν is an approximation of F involving only functions that are described by a *finite* number of parameters, i.e., F^ν is a parametric class. (The sample size n is fixed and we therefore let $\nu \in \mathbb{N}$ index sequences.) We assume that the statistician finds the class F appropriate and, for example, believes it balances over- and underfitting. Consequently, the goal becomes to find a function in (4). The approximation F^ν is introduced for computational reasons and is often selected as close to F as possible, only limited by the computing resources available.

Estimation Algorithm.

Step 0. Set $\nu = 1$.

Step 1. Find $f^\nu \in \varepsilon^\nu\text{-}\operatorname{argmin}_{f \in F^\nu} \frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f) + \pi^\nu(f)$.

Step 2. Replace ν by $\nu + 1$ and go to Step 1.

This seemingly simple algorithm captures a large variety of situations. It constructs a sequence of functions that approximate those in (4) by allowing a tolerance ε^ν that may be larger than ε and by resorting to approximations F^ν and π^ν of the actual quantities F and π. The difficulty in carrying out Step 1 depends on many factors, but since F^ν consists only of functions described by a finite number of parameters it reduces to finite-dimensional optimization for which there are a large number of solvers available. Section 5.2 shows that we often end up with *convex problems.*

The algorithm permits the strategy of initially considering coarse approximations in Step 1 with subsequent refinement. Since iteration number ν has $f^{\nu-1}$ available for warm-staring the computations of f^ν, the amount of computational work required by a solver in Step 1 is often low. In essence, the algorithm can make much progress towards (4) using relatively coarse approximations.

5.1 Theorem (convergence of algorithm). *Suppose that* $x^1, \ldots, x^n \in \mathbb{R}^{d_0}$, $F^\nu, F \subset F^0 \subset$ *usc-fcns*(S) *are closed,* $\psi(x^j, \cdot) : F^0 \to (-\infty, \infty]$ *is continuous for all* j, *and* $\pi, \pi^\nu : F^0 \to \mathbb{R}$ *satisfy* $\pi^\nu(g^\nu) \to \pi(g)$ *whenever* $g^\nu \in F^0 \to g$. *Moreover, let* $\{\varepsilon^\nu \geq 0, \nu \in \mathbb{N}\} \to \varepsilon^\infty$, $\operatorname{Lim} F^\nu = F$, *and* $\{f^\nu, \nu \in \mathbb{N}\}$ *be generated by the Estimation Algorithm.*

(i) If $\varepsilon^\infty \leq \varepsilon$, *then (4) contains every cluster point of* $\{f^\nu, \nu \in \mathbb{N}\}$.

(ii) If $\varepsilon^\infty < \varepsilon$, $F^\nu \subset F$, F^0 *is bounded, and there exists* $g \in F$ *such that* $\psi(x^j, g) < \infty$ *for all* j, *then (4) contains* $f^{\bar{\nu}}$ *for some finite* $\bar{\nu}$.

When $\varepsilon > 0$, item (ii) of the theorem establishes that we obtain an estimate in a *finite* number of iterations of the Estimation Algorithm as long as F^ν approximates F from the "inside." Although not the only possibility, such inner approximations are the primary forms as seen in Section 5.1.

The main technical and practical challenge associated with the Estimation Algorithm is the construction of a parametric class F^ν that set-converges to F. Since F can be a rich class of usc functions, standard approaches (see for example [27, 25, 26]) may fail and we leverage instead a tailored approximation theory for usc-fcns(S).

5.1 Parametric Class of Epi-Splines

Epi-splines is a parametric class that is dense in usc-fcns(S) after a sign change and furnish the building blocks for constructing a parametric class F^ν that approximates F. In essence, an epi-spline on $S \subset \mathbb{R}^d$ is a piecewise polynomial function that is defined in terms of a partition of S consisting of N disjoint open subsets that is dense in S. On each such subset, the epi-spline is a polynomial function. Outside these subsets, the epi-spline is defined by the lower limit of function values making epi-splines lsc; see [34, 30, 33]. Although approximation theory for epi-splines exists for noncompact S, arbitrary partitions, and higher-order polynomials, we here develop the possibilities in the statistical setting for a compact polyhedral $S \subset \mathbb{R}^d$, simplicial complex partitions, and first-degree polynomials.

We denote by $\operatorname{cl} A$ the closure of a set $A \subset \mathbb{R}^d$. A collection $\mathcal{R} = \{R_k\}_{k=1}^N$ of open subsets of S is a *simplicial complex partition* of S if $\operatorname{cl} R_1, \ldots, \operatorname{cl} R_N$ are simplexes[10], $\cup_{k=1}^N \operatorname{cl} R_k = S$,

[10] A *simplex* in $\mathbb{R}^d$ is the convex hull of $d+1$ points $x^0, x^1, \ldots, x^d \in \mathbb{R}^d$, with $x^1 - x^0$, $x^2 - x^0$, $\ldots$, $x^d - x^0$ linearly independent.

and $R_k \cap R_l = \emptyset, k \neq l$. Suppose that $\{\mathcal{R}^\nu = (R_1^\nu, \ldots, R_{N^\nu}^\nu), \nu \in I\!N\}$ is a collection of simplicial complex partition of S with mesh size $\max_{k=1,\ldots,N^\nu} \sup_{x,y \in R_k^\nu} \|x - y\| \to 0$ as $\nu \to \infty$.

A *first-order epi-spline* s on a simplicial complex partition $\mathcal{R} = \{R_k\}_{k=1}^N$ is a real-valued function that on each R_k is affine and that satisfies $\liminf s(x^\nu) = s(x)$ for all $x^\nu \to x$. Let e-spl($\mathcal{R}$) be the collection of all such epi-splines. We deduce from [34, 30] that

$$\bigcup_{\nu \in N} \big\{ f : S \to I\!R \;\big|\; f = -s,\ s \in \text{e-spl}(\mathcal{R}^\nu) \big\} \text{ is dense in } \big(\text{usc-fcns}(S), d\!l \big).$$

In the context of the Estimation Algorithm and Theorem 5.1, this fact underpins several approaches to constructing a parametric class F^ν that set-converges to F. For example, suppose that F is solid[11], then $F^\nu = F \cap \text{e-spl}(\mathcal{R}^\nu) \to F$ as can be established by a standard triangular array argument. One particular class of functions that always will be solid is F_δ^n in Theorem 3.14 provided that it is a subset of a convex F^0. For example, F^0 can be taken to be $\{f \in \text{usc-fcns}(S) \mid f(x) \geq \alpha \ \forall x \in S\}$, which is convex, so this is no real limitation. Consequently, the relaxation of F^n to F_δ^n in Theorem 3.14 not only facilitates consistency of an estimator, it also supports the development of computational methods.

5.2 Examples of Formulations

If F^ν is defined in terms of first-order epi-splines on a partition of $S \subset I\!R^d$ consisting of N^ν open sets, then each function in F^ν is characterized by $N^\nu(d+1)$ parameters. Consequently, Step 1 of the Estimation Algorithm amounts to approximately solving an optimization problem with $N^\nu(d+1)$ variables. The number of variables is independent of the sample size n. The number of open sets N^ν would usually grow with d, but when the growth is slow the number of variables is manageable for modern optimization solvers even for moderately large d.

Among the numerous formulations of the problem in Step 1 of the Estimation Algorithm, we illustrate one based on first-order epi-splines with a simplicial complex partition, which is also used in Section 2.3. Suppose that $c_k^0, c_k^1, \ldots, c_k^d \in I\!R^d$ are the vertexes of the kth simplex of a simplicial complex partition of $S \subset I\!R^d$ with N simplexes. A first-order epi-spline is then fully defined by its height at these vertexes. Let $h_k^i \in I\!R$ be the height at c_k^i, $i = 0, 1, \ldots, d$, $k = 1, \ldots, N$. These $N(d+1)$ variables are to be optimized. (Optimization over such "tent poles" is familiar in ML estimation over log-concave densities, but then they are located at the data points and not according to simplexes as here; see for example [7].) We next give specific expressions for typical objective and constraint functions.

In ML estimation of densities, the loss expressed in terms of the optimization variables becomes

$$-\frac{1}{n} \sum_{j=1}^n \log f(x^j) = -\frac{1}{n} \sum_{j=1}^n \log \sum_{i=0}^d \mu_j^i h_{k_j}^i,$$

[11] A set A is solid if $\text{cl}(\text{int}\, A) = A$.

where k_j is the simplex in which data point x^j is located and the scalars $\{\mu_j^i, i = 0, 1, \ldots, d, j = 1, \ldots, n\}$ can be precomputed by solving $x^j = \sum_{i=0}^d \mu_j^i c_{k_j}^i$. The loss is therefore convex in the optimization variables.

The requirement that functions are nonnegativity is implemented by the constraints $h_k^i \geq 0$ for all $i = 0, 1, \ldots, d,\ k = 1, \ldots, N$, which define a polyhedral feasible set.

The requirement that functions integrate to one is implemented by

$$\int f(x)dx = \frac{1}{d+1} \sum_{k=1}^N \alpha_k \sum_{i=0}^d h_k^i = 1,$$

where α_k is the hyper-volume of the kth simplex.

The requirement that functions should have their argmax covering a given point $x^\star$ is implemented by the constraints

$$\sum_{i=0}^d \eta^i h_{k^\star}^i \geq h_k^{i'} \text{ for all } i' = 0, 1, \ldots, d,\ k = 1, \ldots, N,$$

where $k^\star$ is the simplex in which $x^\star$ is located and the scalars $\{\eta^i, i = 0, 1, \ldots, d\}$ can be precomputed by solving $x^\star = \sum_{i=0}^d \eta^i c_{k^\star}^i$. The constraints form a polyhedral feasible set.

Implementation of continuity, Lipschitz continuity, concavity, and many other conditions also lead to polyhedral feasible sets. Consequently, ML estimation of densities on a compact polyhedral set $S \subset I\!R^d$ under a large variety of constraints can be achieved by optimization of a convex function over a polyhedral feasible sets for which highly efficient solvers are available. A switch to LS regression, would result in a convex quadratic function to minimize, with many of the constraints remaining unchanged. In that case, specialized quadratic optimization solvers apply.

Acknowledgements. This material is based upon work supported in part by ONR Science of Autonomy (N0001417WX01210, N000141712372), DARPA (HR0011834187), and NPS CIMS.

References

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

[2] G. Beer, R. T. Rockafellar, and R. J-B Wets. A characterization of epi-convergence in terms of convergence of level sets. *Proceedings of the American Mathematical Society*, 116(3):753–761, 1992.

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

[4] A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman-Pearson and minmax criteria. Technical Report LA-UR 022951, Los Alamos National Laboratory, Los Alamos, USA, 2002.

[5] D. Casasent and X. Chen. Radial basis function neural networks for nonlinear Fisher discrimination and Neyman-Pearson classification. *Neural Networks*, 16(5):529–535, 2003.

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

[7] 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.

[8] 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.

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

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

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

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

[13] C. R. Doss and J. A. Wellner. Univariate log-concave density estimation with symmetry or modal constraints. *ArXiv e-prints*, November 2019.

[14] L. Dümbgen, R. J. Samworth, and D. Schuhmacher. Approximation by log-concave distributions with applications to regression. *Annals of Statistics*, 39:702–730, 2011.

[15] J. Dupacova and R. J-B Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Annals of Statistics*, 16(4):1517–1549, 1988.

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

[17] S. Fallat, S. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in Markov structures. *Annals of Statistics*, 45(3):1152–1184, 2017.

[18] S. van de Geer and M. Wegkamp. Consistency for the least-squares estimator in nonparametric regression. *Annals of Statistics*, 24(6):2513–2523, 1996.

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

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

[21] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1*, pages 221–233. Univ. of Calif. Press, 1967.

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

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

[24] B. Lavrič. Continuity of monotone functions. *Archivum Mathematicum*, 29(1-2):1–4, 1993.

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

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

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

[28] P. Probst, A.-L. Boulesteix, and B. Bischl. Tunability: Importance of hyperparameters of machine learning algorithms. *J. Machine Learning Research*, 20:1–32, 2019.

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

[30] J. O. Royset. Approximations and solution estimates in optimization. *Mathematical Programming*, 170(2):479–506, 2018.

[31] J. O. Royset. Approximations of semicontinuous functions with applications to stochastic optimization and statistical estimation. *Mathematical Programming*, OnlineFirst, 2019.

[32] J. O. Royset and R. J-B Wets. From data to assessments and decisions: Epi-spline technology. In A. Newman, editor, *INFORMS Tutorials*. INFORMS, Catonsville, 2014.

[33] 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.

[34] 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.

[35] J. O. Royset and R. J-B Wets. Variational theory for optimization under stochastic ambiguity. *SIAM J. Optimization*, 27(2):1118–1149, 2017.

[36] J. O. Royset and R. J-B Wets. Lopsided convergence: an extension and its quantifications. *Mathematical Programming*, 177(1):395–423, 2019.

[37] G. Salinetti and R. J-B Wets. On the convergence in distribution of measurable multifunctions (random sets), normal integrands, stochastic processes and stochastic infima. *Mathematics of Operations Research*, 11(3):385–419, 1986.

[38] G. Salinetti and R. J-B Wets. On the hypo-convergence of probability measures. In *Optimication and Related Fields, Proc., Erice 1984, Lecture Notes in Mathematics 1190*, pages 371–395. Springer, 1986.

[39] S. Sasabuchi, M. Inutsuka, and D. D. S. Kulatunga. A multivariate version of isotonic regression. *Biometrika*, 70(2):465–472, 1983.

[40] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inf. Theory*, 51:3806–3819, 2005.

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

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

[43] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

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

[45] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[46] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

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

[48] A. Waechter. Ipopt interior point optimizer. http://projects.coin-or.org/Ipopt, 2018.

[49] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601, 1949.

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

6 Appendix: Additional Examples

This section discusses existence of solutions of approximation problems for the already well-understood classes of monotone and of log-concave functions. We give proofs passing through the metric space (usc-fcns$(S), d\!\!\!l$) to further illustrate the framework.

6.1 Proposition (existence of monotone LS approximation). *For a box $S \subset I\!R^d$, suppose that $F = \{f \in$ usc-fcns$(S) \mid f$ nondecreasing$\}$ and P is an absolutely continuous distribution on $S \times I\!R$. Then,*

$$\operatorname{argmin}_{f\in F} \int (y - f(x))^2 dP(x, y) \neq \emptyset.$$

Proof. By Proposition 4.3, F is closed. Suppose that $f^n \in F \to f$. Let $D = \{x \in \operatorname{int} S \mid f$ is discontinuous at x In view of Propositions 3.2(iv) and 3.1, $f^n(x) \to f(x)$ for all $x \in \operatorname{int} S \setminus D$. Thus, $(y - f^n(x))^2 \to (y - f(x))^2$ for all such x and all $y \in I\!R$. By [24], D has Lebesgue measure zero and the same holds for $S \setminus \operatorname{int} S$. Then, by Fatou's Lemma, $\liminf \int (y - f^n(x))^2 dP(x, y) \geq \int (y - f(x))^2 dP(x, y)$ and $f \mapsto \int (y - f(x))^2 dP(x, y)$ is lsc on F. Its lower level-sets are compact at every finite level (cf. the argument in the proof of Corollary 3.7) and the conclusion follows. □

The next result is in [14], but we provide a proof with some novel elements: the log-likelihood criterion function is shown to be lsc on the enlarged class of log-concave functions that integrate to values in $[0, 1]$.

6.2 Proposition (existence of log-concave ML estimator). *Suppose that $F = \{f \in$ usc-fcns$(I\!R^d) \mid f \geq 0,$ log-concave$\}$. Then, for any probability distribution P on $I\!R^d$,*

$$\operatorname{argmin}_{f\in F} \Big\{ \int -\log f(x) dP(x) \;\Big|\; \int f(x)dx = 1 \Big\} \neq \emptyset$$

if and only if

$$\int \|x\| dP(x) < \infty \text{ and } P(H) < 1 \text{ for all hyperplane } H \subset I\!R^d.$$

Proof. For $f^n \in F \to f$, Proposition 4.1(ii) establishes that f is log-concave. Moreover, $f^n(x) \to f(x)$ for all $x \in \operatorname{int}\{x \in I\!R^d \mid f(x) > 0\}$ and also when $f(x) = 0$ by Propositions 3.1 and 3.2. The subset of $I\!R^d$ that fails outside both of these cases has Lebesgue measure zero so $f^n(x) \to f(x)$ for Lebesgue-a.e. $x \in I\!R^d$. Fatou's Lemma gives that $\liminf \int f^n(x)dx \geq \int f(x)dx$. Thus, $F_{\leq} = \{f \in F \mid \int f(x)dx \leq 1\}$ is closed and actually compact because all functions in F are nonnegative.

We show that $\varphi(f) = \int -\log f(x) dP(x)$ is lsc as a function on $(F_{\leq}, d\!\!\!l)$. Let $f^n \in F_{\leq} \to f$. We consider two cases: a) $\int f(x)dx = \gamma > 0$. Then, $\gamma^{-1} f$ is a log-concave density and by [8, Lem. 1] there are $\xi_0 \in I\!R$ and $\xi_1 \in (0, \infty)$ such that $f(x) \leq \exp(\xi_0 - \xi_1 \|x\|)$ for all $x \in I\!R^d$. Let $\varepsilon = \sup_{x\in R^d} f(x)/4$, which then must be positive, and $\rho \in (2\varepsilon, \infty)$ such that $f(x) \leq \varepsilon$ for

$\|x\|_2 \geq \rho/2$. (Here, we adopt the Euclidean norm, with the correspond balls denoted by $\mathbb{B}_2(x,\delta)$, to simplify a reference to [29].) Hypo-convergence is locally uniform in the following sense [29, Thm. 4.10]: there is $\bar{n}$ such that for $n \geq \bar{n}$,

$$\begin{aligned} \operatorname{hypo} f^n \cap \mathbb{B}_2(0,\rho) &\subset \operatorname{hypo} f + \mathbb{B}_2(0,\varepsilon) \\ \operatorname{hypo} f \cap \mathbb{B}_2(0,\rho) &\subset \operatorname{hypo} f^n + \mathbb{B}_2(0,\varepsilon). \end{aligned}$$

Take $(x, f^n(x))$ with $\|x\|_2 = \rho$. If $f^n(x) > \rho$, then $(x,\rho) \in \operatorname{hypo} f^n \cap \mathbb{B}_2(0,\rho)$ and there exists $(y,\beta) \in \operatorname{hypo} f$ such that $\|x-y\|_2 \leq \varepsilon$ and $|\rho - \beta| \leq \varepsilon$. Thus, $f(y) \geq \beta \geq \rho - \varepsilon > \varepsilon$. However, $f(y) \leq \varepsilon$ because $\|y\|_2 \geq \rho/2$ and we have reached a contradiction. Thus, $f^n(x) \leq \rho$, $(x, f^n(x)) \in \operatorname{hypo} f^n \cap \mathbb{B}_2(0,\rho)$, and there is $(y,\beta) \in \operatorname{hypo} f$ such that $\|x-y\|_2 \leq \varepsilon$ and $|f^n(x) - \beta| \leq \varepsilon$. This leads to $f^n(x) \leq \beta + \varepsilon \leq f(y) + \varepsilon \leq 2\varepsilon$ for all $n \geq \bar{n}$. The choice of ρ ensures that $\bar{x} \in \operatorname{argmax}_{x \in \mathbb{R}^d} f(x)$ with $\|\bar{x}\|_2 \leq \rho/2$ exists. By (3), there is $x^n \to \bar{x}$ such that $f^n(x^n) \to f(\bar{x}) = 4\varepsilon$. Thus, for some $n^* \geq \bar{n}$, $\|x^n\|_2 \leq 3\rho/4$ and $f^n(x^n) \geq 3\varepsilon$ for all $n \geq n^*$. Since we also have $f^n(x) \leq 2\varepsilon$ for $\|x\|_2 = \rho$, $\operatorname{argmax}_{x \in \mathbb{R}^d} f^n(x) \subset \mathbb{B}_2(0, 3\rho/4)$ for all $n \geq n^*$. By [29, Thm. 7.31], this implies that $\sup_{x \in \mathbb{R}^d} f^n(x) \to \sup_{x \in \mathbb{R}^d} f(x)$. Consequently, for sufficiently large n, $\int -\log f^n(x) dP(x) \geq \int -\log[2 \sup_{\bar{x} \in \mathbb{R}^d} f(\bar{x})] dP(x) > -\infty$, which then furnishes an integrable lower for application of Fatou's lemma: $\operatorname{liminf} \int -\log f^n(x) dP(x) \geq \int \operatorname{liminf}[-\log f^n(x)] dP(x)$. Since $\operatorname{liminf} -\log f^n(x) \geq -\log f(x)$ for all $x \in \mathbb{R}^d$ by (3), we conclude that φ is lsc at points $f \in F_{\leq}$ with $\int f(x)dx > 0$. This fact holds for any P.

Next, we consider b) $\int f(x)dx = 0$ and now it becomes essential to limit the scope to P with the stated properties. Let $D = \{x \in \mathbb{R}^d \mid f(x) > 0\}$, which then has Lebesgue measure zero (because $\int f(x)dx = 0$) and $\operatorname{int} D = \emptyset$. Since D is also convex by the log-concavity of f, it lies in an affine subspace of $\mathbb{R}^d$ of dimension less than d, i.e., D is a subset of some hyperplane $H \subset \mathbb{R}^d$. Consequently, the first term of

$$\varphi(f) = \int_{x \notin D} -\log f(x) dP(x) + \int_{x \in D} -\log f(x) dP(x)$$

integrates to ∞ in view of the assumption on P. The convention $\infty - \alpha = \infty$ for any $\alpha \in \overline{\mathbb{R}}$ implies that $\varphi(f) = \infty$ regardless of the value of the second term. It remains to show that $\varphi(f^n) \to \infty$ when $f^n \in F_{\leq} \to f$. Since $\varphi(f^n) = \infty$ when $\int f^n(x)dx = 0$ as just argued, we assume without loss of generality that $\int f^n(x)dx > 0$ for all n. In fact, those integrals can be assumed to be one because, with $\int f^n(x)dx = \gamma^n$, $\int -\log f^n(x) dP(x) = -\log \gamma^n + \int -\log(f^n(x)/\gamma^n) dP(x) \to \infty$ when the last term tends to ∞.

Each $s^n = \sup_{x \in \mathbb{R}^d} f^n(x)$, $n \in \mathbb{N}$, is finite (cf. [8, Lem. 1]), but the sequence could be unbounded. If $\sup_{n \in \mathbb{N}} s^n$ is also finite, then

$$\begin{aligned} \varphi(f^n) = &\int_{f(x)=0, f^n(x) \leq 1} -\log f^n(x) dP(x) + \int_{f(x)>0, f^n(x) \leq 1} -\log f^n(x) dP(x) \\ &+ \int_{f^n(x)>1} -\log f^n(x) dP(x) \to \infty; \end{aligned}$$

the first term tends to ∞ because $f^n(x) \to 0$ when $f(x) = 0$ by Proposition 3.2(i) and the last term is bounded from below uniformly in n. Hence, suppose that $s^n \to \infty$. For $\eta > 0$, $\tau^n = \log s^n$, and $\sigma^n = \exp(-\eta\tau^n)$,

$$\begin{aligned}\int -\log f^n(x) dP(x) &\geq \eta\tau^n P(I\!R^d \setminus \mathrm{lev}_{\geq \sigma^n} f^n) - \tau^n P(\mathrm{lev}_{\geq \sigma^n} f^n) \\ &= (\eta+1)\tau^n \Big(\frac{\eta}{\eta+1} - P(\mathrm{lev}_{\geq \sigma^n} f^n) \Big).\end{aligned}$$

By [14, Lem. 4.1], the Lebesgue measure of $\mathrm{lev}_{\geq \sigma^n} f^n$ is no greater than

$$(1+\eta)^d (\tau^n)^d \exp(-\tau^n) / \int_0^{(1+\eta)\tau^n} t^d \exp(-t) dt \to 0$$

as s^n (and τ^n) tends to infinity for any given $\eta > 0$. Moreover, [14, Lem. 2.1] establishes that $P(\mathrm{lev}_{\geq \sigma^n} f^n) < \eta/(\eta+1)$ when the Lebesgue measure of $\mathrm{lev}_{\geq \sigma^n} f^n$ is sufficiently low and η sufficiently high. (This fact relies critically on the assumption on P.) Thus, $\int -\log f^n(x) dP(x) \to \infty$ when $s^n \to \infty$ and φ is lsc (in fact continuous) at f when $\int f(x)dx = 0$.

In summary, we have shown that φ is lsc on the compact set $F_{\leq}$. Thus, there exists $f^\star \in \mathrm{argmin}_{f \in F_{\leq}} \varphi(f)$. Trivially, there is $f \in F_{\leq}$ with finite $\varphi(f)$, which implies that $\varphi(f^\star) < \infty$ and, as argued above, $\int f^\star(x)dx = \gamma > 0$. Since $\varphi(f^\star/\gamma) \leq \varphi(f^\star)$, $f^\star/\gamma \in \mathrm{argmin}_{f \in F}\{\varphi(f) \mid \int f(x)dx = 1\}$.

For the necessity of the conditions on P we refer to [14]. □

7 Appendix: Intermediate Results and Proofs

This section includes proofs of all the results in the paper.

Proof of Proposition 2.1. By Proposition 3.2(v), F is equi-usc at all $x \in I\!R^d$. Proposition 4.8(i) ensures that the integral constraint is closed. Theorem 3.16 as well as Propositions 4.4 and 4.5 establish that the other constraints are closed too. Consequently, F is compact. Corollary 3.8 applies and confirms (i). Corollary 3.12 and the discussion immediately after it establish (ii). When $f^0 \in F$, then every cluster point of $\{\hat{f}^n, n \in I\!N\}$ must deviate from f^0 at most on set of Lebesgue measure zero. For Lipschitz continuous functions this means that the functions must be identical and (iii) holds. □

Proof of Proposition 3.2. When $f^n \to f$, it suffices by [29, Thm. 7.10] to establish that $f^n(\bar{x}) \to f(\bar{x})$. In view of (3), (i) is trivial. Items (ii,iii) follow by [29, Thm. 7.17]. For (iv), we only prove the nondecreasing case as a nearly identical argument establishes the conclusion for nonincreasing functions. Let $\varepsilon > 0$. Since $\bar{x} \in \mathrm{int}\, S$ and f is continuous at $\bar{x}$, there exist $\bar{y} \in S$, with $\bar{y}_i < \bar{x}_i$ for $i = 1, \ldots, d$, and $f(\bar{y}) \geq f(\bar{x}) - \varepsilon$. Moreover, for some $x^n \in S \to \bar{y}$, $f^n(x^n) \to f(\bar{y})$

by (3). Since $x^n \leq \bar{x}$ for sufficiently large n, $\operatorname{liminf} f^n(\bar{x}) \geq \operatorname{liminf} f^n(x^n) = f(\bar{y}) \geq f(\bar{x}) - \varepsilon$. Since ε is arbitrary, the conclusion follows because $\operatorname{limsup} f^n(\bar{x}) \leq f(\bar{x})$ already by (3). For (v), consider the definition of equi-usc. The Lipschitz condition ensures that there is $\delta \in (0, \infty)$ with $f^n(x) \leq f^n(\bar{x}) + \kappa \|x - \bar{x}\|$ for all $n \in \mathbb{N}$ and $x \in \mathbb{B}(\bar{x}, \delta)$. Let $\varepsilon > 0$. If $\kappa = 0$, then set $\delta' = \delta$. Otherwise, set $\delta' = \min\{\varepsilon/\kappa, \delta\}$. In either case, $\sup_{x \in B(\bar{x}, \delta')} f^n(x) \leq f^n(\bar{x}) + \kappa\delta' \leq f^n(\bar{x}) + \varepsilon$. □

Proof of Proposition 3.3. In view of Proposition 3.1, the result follows directly from applications of the Dominated Convergence Theorem. □

Proof of Theorem 3.4. A trivial generalization of Fatou's Lemma shows that $\int \psi(x, \cdot) dP^0(x)$ is lsc on F (see for example [12, Appendix]). Since for all $f \in F$, $\int \psi(x, f) dP^0(x)$ and $\pi(f)$ exceed $-\infty$, $\int \psi(x, \cdot) dP^0(x) + \pi$ is lsc on F. All lsc functions defined on a compact set attain their infima. □

Proof of Corollary 3.5. The function $n^{-1} \sum_{j=1}^n \psi(x^j, \cdot) + \pi$ is lsc on F because each term in the sum involves a lsc function that is never $-\infty$. All lsc functions defined on a compact set attain their infima. □

Proof of Corollary 3.6. By Propositions 4.1(i) and 4.5 as well as Theorem 3.16, F is closed. It is also bounded; see the remark after Corollary 3.5. Since $g > -\infty$ and $x^j \in \operatorname{int} S$, it is also equi-usc at x^j, $j = 1, \ldots, n$, by Proposition 3.2(ii). Thus, in view of Proposition 3.1 $f \mapsto \max\{0, 1 - y^j f(x^j)\}$ is continuous on F for all j and the conclusion follows by Corollary 3.5. □

Proof of Corollary 3.7. Let $\varphi(f) = n^{-1} \sum_{j=1}^n \max\{0, 1 - f(x^j)\}$, $f \in F$, and $f^n \in F \to f$. By (3), $\operatorname{limsup} f^n(x^j) \leq f(x^j)$ and $\operatorname{liminf}(\max\{0, 1 - f^n(x^j)\}) \geq \max\{0, 1 - f(x^j)\}$ for all $j = 1, \ldots, n$, which implies that φ is lsc on F. Since Z^i is open, $\operatorname{liminf}(\sup_{x \in Z^i} f^n(x)) \geq \sup_{x \in Z^i} f(x)$ by [29, Prop. 7.29]. Thus, $F^0 = \{f \in F \mid \sup_{x \in Z^i} f(x) \leq 0, i = 1, \ldots, m\}$ is closed. For $g \in F$, suppose that $\{f^n \in F, n \in \mathbb{N}\}$ is such that $d\!l(f^n, g) \to \infty$. Then, $\operatorname{hypo} f^n$ set-converges to $\emptyset$, which implies $f^n(x^j) \to -\infty$ for all j and $\varphi(f^n) \to \infty$. Consequently, $\{f \in F \mid \varphi(f) \leq \alpha\}$ is bounded for $\alpha \in \mathbb{R}$. Since it is also closed by virtue of φ being lsc, these level-sets are actually compact. A lsc function with compact level-set attains it infimum. □

Proof of Corollary 3.8. Let $f^n \in F \to f$. By (3), $\operatorname{limsup} f^n(x^j) \leq f(x^j)$ for all j so $\operatorname{liminf} - \log f^n(x^j) \geq - \log f(x^j)$. Thus, $f \mapsto -n^{-1} \sum_{j=1}^n \log f(x^j)$ is lsc on F. The conclusion then follows by Corollary 3.5. □

Proof of Corollary 3.9. In view of Proposition 3.1, $f \mapsto \sum_{j=1}^n (y^j - f(x^j))^2$ is continuous on F. An argument similar to the one in the proof of Corollary 3.7 yields that this function has compact level-sets. □

The proof of Theorem 3.10 relies on an lsc-LLN, essentially in [1, 23], that ensures almost sure epi-convergence of empirical processes indexed on a polish space. For completeness, we include the statement as well as a new proof, which is simpler than that in [1]. It follows the arguments

in [23] for ergodic processes, but takes advantage of the present iid setting. The statement is made slightly more general than needed without complication.

7.1 Proposition (lsc-LLN). *Suppose that (Y, d_Y) is a complete separable (polish) metric space, $(\Xi, \mathcal{A}, P)$ is a complete*[12] *probability space, and $\psi : \Xi \times Y \to \overline{I\!R}$ is a locally inf-integrable random lsc function*[13]*. If $\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \dots$ is a sequence of iid random elements that take values in Ξ with distribution P, then almost surely*

$$\frac{1}{n}\sum_{j=1}^{n} \psi(\boldsymbol{\xi}^j, \cdot) \text{ epi-converges } I\!E[\psi(\boldsymbol{\xi}^1, \cdot)],$$

which is equivalent to having for all $y \in Y$,

$$\forall y^n \to y, \operatorname{liminf} \frac{1}{n}\sum_{j=1}^{n} \psi(\boldsymbol{\xi}^j, y^n) \geq I\!E[\psi(\boldsymbol{\xi}^1, y)]$$

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

Proof. A slight generalization of Fatou's Lemma (see [12, Appendix]) ensures that $I\!E[\psi(\boldsymbol{\xi}^1, \cdot)]$ is lsc. Let $\bar{D} \subset Y \times \overline{I\!R}$ be a countable dense subset of the epigraph $\operatorname{epi} I\!E[\psi(\boldsymbol{\xi}^1, \cdot)]$, with $\operatorname{epi} h = \{(y, y_0) \in Y \times I\!R \mid h(y) \leq y_0\}$, which may be empty. 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

$$\pi_{y,r}(\xi) = \inf_{y' \in I\!B^o(y,r)} \psi(\xi, y') \text{ if } r > 0 \text{ and } \pi_{y,0}(\xi) = \psi(\xi, y),$$

where $I\!B^o(y, r) = \{y' \in Y \mid d_Y(y', y) < r\}$. By Theorem 3.4 in [23], 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 is a closed neighborhood V_y of y and $r_y \in (0, \infty)$ such that

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

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 [16, Thms. 7.1,7.2]) ensures that

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

[12] In view of [23], the result (but not our proof) holds without completeness.

[13] The definitions of Section 3.2 carry over to the more general context here.

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

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

We proceed by establishing the liminf and limsup conditions of the theorem. 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 $(\xi^1, \xi^2, \dots) \in \Xi_0^\infty$,

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

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, $I\!E[\pi_{z^k, r^k}] \to I\!E[\psi(\boldsymbol{\xi}^1, y)]$. We have establish that for $(\xi^1, \xi^2, \dots) \in \Xi_0^\infty$, $\liminf n^{-1} \sum_{j=1}^n \psi(\xi^j, y^n) \geq I\!E[\psi(\boldsymbol{\xi}^1, y)]$.

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

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

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

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

Proof of Theorem 3.10. If F is empty, the results hold trivially. Suppose that F is nonempty. By [29, Prop. 4.45, Thm. 7.58], $(\text{usc-fcns}(S), dl)$ is a complete separable metric space. By virtue of being a closed subset, F forms another complete separable metric space (F, dl). Let $\varphi^n(f) = n^{-1} \sum_{j=1}^n \psi(X^j, f) + \pi^n(f)$ and $\varphi(f) = I\!E[\psi(X^1, f)]$, $f \in F$. Proposition 7.1 applied with this metric space establishes that $n^{-1} \sum_{j=1}^n \psi(X^j, \cdot)$ epi-converges to φ a.s. Moreover, for all $f^n \in F \to f$,

$$\liminf \varphi^n(f^n) \geq \liminf \frac{1}{n}\sum_{j=1}^{n} \psi(X^j, f^n) \geq \varphi(f) \text{ a.s.}$$

Also, there exists $f^n \in F \to f$ such that

$$\operatorname{limsup} \varphi^n(f^n) \leq \operatorname{limsup} \frac{1}{n} \sum_{j=1}^{n} \psi(X^j, f^n) + \operatorname{limsup} \pi^n(f^n) \leq \varphi(f) \text{ a.s.}$$

We have established that φ^n epi-converges to φ a.s. If φ is improper, which in this case means that $\varphi(f) = \infty$ for all $f \in F$, then item (i) holds trivially because the right-hand side of the inclusion is the whole of F. If φ is proper, then φ^n is also proper and [30, Prop. 2.1] applies, which establishes (i).

The additional assumptions in item (ii) imply that both φ and φ^n are proper, and also that φ^n epi-converges tightly φ because then F is compact. Thus, [36, Thm. 3.8] applies and item (ii) is established. □

Proof of Corollary 3.11. We deduce from the proof of Corollary 3.6 that F is compact and equi-usc at all $x \in \mathbb{R}^d$. This implies that for all $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$, $f \mapsto \max\{0, 1 - yf(x)\}$ is continuous on F. Suppose that $f^n \to f$ and $x^n \to x$. By (3), $\operatorname{limsup} f^n(x^n) \leq f(x)$. Thus, $(x, f) \mapsto f(x)$ is usc on $\mathbb{R}^d \times F$. From this we conclude that $((x, y), f) \mapsto \max\{0, 1 - yf(x)\}$ is measurable and a random lsc function. It is trivally locally inf-integrable by virtue of being nonnegative. Theorem 3.10(i) therefore applies and a cluster point $f^\star$ of $\{\hat{f}^n, n \in \mathbb{N}\}$, of which one exits due to compactness of F, must satisfy the first conclusion a.s. The second conclusion follows by an application of [29, Prop. 7.7]. □

Proof of Corollary 3.12. Since F consists of nonnegative functions, it is bounded and in fact compact since closed. Thus, $\{\hat{f}^n, n \in \mathbb{N}\}$ must have at least one cluster point. Next, we show that $\psi : S \times F \to \overline{\mathbb{R}}$ given by $\psi(x, f) = -\log f(x)$ is a random lsc function. Suppose that $f^n \in F \to f$ and $x^n \in S \to x$, then $\operatorname{limsup} f^n(x^n) \leq f(x)$ and also $\operatorname{liminf} -\log f^n(x^n) \geq -\log f(x)$, which implies that ψ is lsc. Measurability then follows directly from the fact that lower level-sets of lsc functions are closed. Theorem 3.10(i) therefore applies and a cluster point $f^\star$ of $\{\hat{f}^n, n \in \mathbb{N}\}$ must satisfy a.s.

$$f^\star \in \operatorname{argmin}_{f \in F} \mathbb{E}[-\log f(X^1)] \subset \operatorname{argmin}_{f \in F} \mathbb{E}[\log f^0(X^1)] - \mathbb{E}[\log f(X^1)].$$

The inclusion holds even if $\mathbb{E}[\log f^0(X^1)]$ equals $-\infty$ or ∞. The last conclusion of the theorem follows directly from the properties of the Kullback-Leibler divergence. □

Proof of Corollary 3.13. From the proof of Corollary 3.9, we deduce that $f \mapsto (y - f(x))^2$ is continuous for any $(x, y) \in S \times \mathbb{R}$. Moreover, if $f^n \in F \to f$ and $x^n \in S \to x$, then $\operatorname{limsup} f^n(x^n) \leq f(x)$ by (3). Thus, the mapping $(x, f) \mapsto f(x)$ on $S \times F$ is usc and thus measurable. We therefore have that $((x, y), f) \mapsto (y - f(x))^2$ is measurable too as a function on $S \times \mathbb{R} \times F$ and also a random lsc function. It is trivially locally inf-integrable by its nonnegativity. Theorem 3.10(i) therefore applies and a cluster point $f^\star$ of $\{\hat{f}^n, n \in \mathbb{N}\}$ must satisfy a.s.

$$f^\star \in \operatorname{argmin}_{f \in F} \mathbb{E}\big[(Y^1 - f(X^1)^2\big] = \operatorname{argmin}_{f \in F} L_P^2(f^0, f)$$

because $\mathbb{E}[Z^1] = 0$ and X^1 and Z^1 are independent; the finite variance of Z^1 prevents $\mathbb{E}\big[(Y^1 - f(X^1)^2\big]$ from being ∞ when $L_P^2(f^0, f)$ is finite. The existence of a cluster point is realized as follows. Let $\varphi(f) = \mathbb{E}[(Y^1 - f(X^1)^2]$, $f \in F$. If $\{f^n \in F, n \in \mathbb{N}\}$ satisfies $d\!\!l(f^n, g) \to \infty$ for some $g \in F$, then hypo f^n set-converges to $\emptyset$ and $f^n(x) \to -\infty$ for all $x \in S$. Thus, $\varphi(f^n) = \mathbb{E}[(f^0(X^1) - f^n(X^1)^2] + \mathbb{E}[(Z^1)^2] \to \infty$ since f^0 is real-valued. This implies that $\{f \in F \mid \varphi(f) \leq \alpha\}$ is bounded for all $\alpha \in \mathbb{R}$ and contains OutLim$\{f \in F \mid \varphi^n(f) \leq \alpha\}$ for any sequence of functions $\{\varphi^n : F \to \overline{\mathbb{R}}, n \in \mathbb{N}\}$ epi-converging to φ [2, Thm. 3.1] and also limsup$(\inf_{f\in F} \varphi^n(f)) \leq \inf_{f\in F} \varphi(f)$ [36, Thm. 3.8]. Under the assumption that $\inf_{f\in F} \varphi(f) < \infty$, we therefore have that for some $\bar{n}$, $\{\varepsilon^n\text{-}\operatorname{argmin}_{f\in F} \varphi^n(f), n \geq \bar{n}\}$ is bounded. Applying these facts to the (random) function defined by $\varphi^n(f) = n^{-1} \sum_{j=1}^n (Y^j - f(X^j))^2$, which epi-converges to φ almost surely (cf. Theorem 3.10), establishes that $\{\hat{f}^n, n \in \mathbb{N}\}$ is bounded almost surely and therefore must have a cluster point. The final conclusion follows directly from the properties of the L_P^2 distance. □

Proof of Theorem 3.14. Following the arguments in the proof of Theorem 3.10, we established that $f \mapsto \varphi^n(f) = n^{-1} \sum_{j=1}^n \psi(X^j, \cdot) + \pi^n$ epi-converges to $f \mapsto \varphi(f) = \mathbb{E}[\psi(X^1, \cdot)]$ a.s. as functions on $(F, d\!\!l)$. Next, suppose that

$$f^\star \in \text{OutLim}\left(\varepsilon^n\text{-}\operatorname{argmin}_{f\in F_\delta^n} \varphi^n(f)\right).$$

Then there exist a subsequence $\{n_k, k \in \mathbb{N}\}$ and

$$f^k \in \varepsilon^{n_k}\text{-}\operatorname{argmin}_{f\in F_\delta^{n_k}} \varphi^{n_k}(f) \to f^\star.$$

The continuity of the point-to-set distance and the fact that $\text{dist}(f^k, F^{n_k}) \leq \delta$ for all k implies that $\text{dist}(f^\star, \text{Lim}\, F^n) \leq \delta$, i.e., $f^\star \in F_\delta^\infty$. Thus, it only remains to show that $\varphi(f^\star) \leq \inf_{f\in \text{Lim}\, F^n} \varphi(f)$. Let $g^\star \in \operatorname{argmin}_{f\in \text{Lim}\, F^n} \varphi(f)$. Then, because φ^n epi-converges to φ, there exists $g^n \in F \to g^\star$ such that

$$\text{limsup}\, \varphi^n(g^n) \leq \varphi(g^\star).$$

Since $g^\star \in \text{Lim}\, F^n$, there is $\bar{n} \in \mathbb{N}$ such that $\text{dist}(g^n, F^n) \leq \delta$ for all $n \geq \bar{n}$. Consequently, leveraging the epi-convergence property and the above facts,

$$\begin{aligned}\varphi(f^\star) &\leq \text{liminf}\, \varphi^{n_k}(f^k) \leq \text{liminf}\left(\inf_{f\in F_\delta^{n_k}} \varphi^{n_k}(f) + \varepsilon^{n_k}\right) \\ &\leq \text{limsup}\, \varphi^{n_k}(g^{n_k}) \leq \varphi(g^\star) = \inf_{f\in \text{Lim}\, F^n} \varphi(f).\end{aligned}$$

The first conclusion is established. The second conclusion is immediate after realizing that $F_\delta^\infty = F$ when $\text{Lim}\, F^n = F$. □

Proof of Corollary 3.15. The arguments of Corollary 3.12 in conjunction with Theorem 3.14 yield $f^\star \in F_\delta^\infty$ and $K(f^0; f^\star) \leq \inf_{g\in \text{Lim}\, F^n} K(f^0; g)$. Since $\text{Lim}\, F^n \subset F$ consists only of densities and $f^0 \in \text{Lim}\, F^n$, the right-hand side in this inequality is zero and the conclusion follows. □

Proof of Theorem 3.16. The assertions about $\hat{m}^n$ and $\hat{h}^n$ are essentially in [30, Prop. 2.1], with an extension to improper functions following straightforwardly. The conclusion about $\hat{l}^n$ holds by [29, Prop. 7.7]. □

7.2 Lemma (hypo-convergence under composition). *For a continuous nondecreasing function $h_0 : \mathbb{R} \to \overline{\mathbb{R}}$, let $h : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ have $h(y) = h_0(y)$ if $y \in \mathbb{R}$, $h(-\infty) = \inf_{\bar{y}\in R} h_0(\bar{y})$, and $h(\infty) = \sup_{\bar{y}\in R} h_0(\bar{y})$. If $g^n : S \to \overline{\mathbb{R}}$ hypo-converges to $g : S \to \overline{\mathbb{R}}$, then $h \circ g^n$ hypo-converges to $h \circ g$.*

Proof. Suppose that $x^n \in S \to x$, which implies that $\operatorname{limsup} g^n(x^n) \leq g(x)$. Fix n and let $\varepsilon > 0$. Suppose that $\xi^n = \sup_{m\geq n} h(g^m(x^m)) \in \mathbb{R}$. Then, there exists $\bar{m} \geq n$ such that $\xi^n \leq h(g^{\bar{m}}(x^{\bar{m}}))+\varepsilon \leq h(\sup_{m\geq n} g^m(x^m))+\varepsilon$, the last inequality holds because h is nondecreasing. Since ε is arbitrary, $\xi^n \leq h(\sup_{m\geq n} g^m(x^m))$. A similar argument leads to the same inequality if $\xi^n = \infty$ and, trivially, also when $\xi^n = -\infty$. Since the inequality holds for all n, it follows by the continuity of h that

$$\begin{aligned}\operatorname{limsup} h(g^n(x^n)) &= \lim_{n\to\infty} \big(\sup_{m\geq n} h(g^m(x^m))\big) \leq \lim_{n\to\infty} h\big(\sup_{m\geq n} g^m(x^m)\big)\\ &= h\big(\lim_{n\to\infty} (\sup_{m\geq n} g^m(x^m))\big) = h(\operatorname{limsup} g^n(x^n)) \leq h(g(x)).\end{aligned}$$

For any $x \in S$, there exists $x^n \in S \to x$ with $g^n(x^n) \to g(x)$. Since h is continuous, this implies $h(g^n(x^n)) \to h(g(x))$ and the conclusion follows. □

Proof of Proposition 4.1. The first claim follows by [29, Prop. 4.15]. Since $-f^n, -f$ are proper, lsc, and convex, it follows by [29, Thm. 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 $\|v\|_2 \leq \kappa$.

For part (ii), Lemma 7.2, with h defined by $h(y) = \log y$ if $y \in (0, \infty)$, $h(y) = -\infty$ if $y = [-\infty, 0]$, and $h(y) = \infty$ if $y = \infty$, yields that $h \circ f^n$ hypo-converges to $h \circ f$. Since $h \circ f^n$ is concave, it follows by [29, Prop. 4.15] that $h \circ f$ is concave too.

For part (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 $\varepsilon > 0$ and suppose that $f(z) < \infty$, $f(x) > -\infty$, and $f(y) > -\infty$. 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 $n \geq \bar{n}$

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

Since $\varepsilon > 0$ is arbitrary, $f(z) \leq \lambda f(x) + (1-\lambda) f(y)$. A similar argument leads to the same conclusion when $f(z) = \infty$, $f(x) = -\infty$, and/or $f(y) = -\infty$.

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 $y \in S \backslash \operatorname{int} S$. Then, there exists $y^n \in \operatorname{int} S \to y$ with $f(\lambda x + (1-\lambda) y^n) \leq \lambda f(x) + (1-\lambda) f(y^n)$ because S must be convex. 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 $\lambda f(x) + (1-\lambda) f(y)$ by the usc of f. A similar argument holds in the other cases. Thus, f is convex. □

Proof of Proposition 4.2. By [29, Thm. 7.6], either $\operatorname{hypo} g^n$ set-converges to $\emptyset$ or there exist $g \in \text{usc-fcns}(S)$ and a subsequence $\{n_k, k \in I\!N\}$ such that $g^{n_k} \to g$. In the second case, $h \circ g^{n_k}$ hypo-converges to $h \circ g$ by Lemma 7.2. Since a hypo-limit is unique, $f = h \circ g$. In the first case, for all $x \in S$, $g^n(x) \to -\infty$ so that $h \circ g^n(x) \to f(x) = h(-\infty) = h \circ g(x)$, when $g(x) = -\infty$ for all $x \in S$. The conclusions then follow by 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 $\|z - y\| \leq \delta$. Since $y \in \operatorname{int} S$, z can be takes such that $z_i > y_i$ for $i = 1, \ldots, d$ and $z \in \operatorname{int} S$. By 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. By the nondecreasing property,

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

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 $z \in I\!R^d$ by setting $z_i = \min\{\beta_i, y_i + \delta\}$. Let $\bar{n}$ be such that $x_i^n \leq x_i + \delta$ for all $i = 1, \ldots, d$ 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 $\|z - x\| \leq \delta$. Since $x \in \operatorname{int} S$, z can be takes such that $z_i < x_i$ for $i = 1, \ldots, d$ 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 $i = 1, \ldots, d$ 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 $x, y \in S$, with $f(x)$ and $f(y)$ finite, 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 when finite. If f is not finite on S, then it cannot be Lipschitz continuous. □

Proof of Proposition 4.5. Let $x \in S$ and observe that $g(x) \leq \operatorname{limsup} f^n(x) \leq f(x)$ by (3), 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. □

Proof of Proposition 4.6. Since the collection of functions is equi-usc, hypo-convergence implies pointwise convergence (Proposition 3.1) and the conclusion follows immediately. □

Proof of Proposition 4.7. Let $\varepsilon > 0$ and $f^n \in F \to f$. First, suppose that $\sup_{x\in S} g(f(x)) \in I\!R$. Then, there exists $\bar{x} \in S$ such that $g(f(\bar{x})) \geq \sup_{x\in S} g(f(x)) - \varepsilon$. By (3), there is $x^n \in S \to \bar{x}$ such that $f^n(x^n) \to f(\bar{x})$. Since g is lsc,

$$\operatorname{liminf}\left(\sup_{x\in S} g(f^n(x))\right) \geq \operatorname{liminf} g(f^n(x^n)) \geq g(f(\bar{x})) \geq \sup_{x\in S} g(f(x)) - \varepsilon.$$

Second, suppose that $\sup_{x\in S} g(f(x)) = \infty$. Then, there exists $\bar{x} \in S$ such that $g(f(\bar{x})) \geq 1/\varepsilon$. Again, there is $x^n \in S \to \bar{x}$ such that $f^n(x^n) \to f(\bar{x})$ and

$$\operatorname{liminf}\left(\sup_{x\in S} g(f^n(x))\right) \geq \operatorname{liminf} g(f^n(x^n)) \geq g(f(\bar{x})) \geq 1/\varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, we have established that $\operatorname{liminf}\left(\sup_{x\in S} g(f^n(x))\right) \geq \sup_{x\in S} g(f(x))$; it trivially holds when $\sup_{x\in S} g(f(x)) = -\infty$. □

Proof of Proposition 4.8. Since the collection of functions is equi-usc at Lebesgue-a.e. $x \in S$, hypo-convergence implies pointwise convergence at Lebesgue-a.e. $x \in S$ by Proposition 3.1. The conclusions follow directly from an application of the Dominated Convergence Theorem. □

Proof of Proposition 4.9. Since $C \subset C^n$, $F \subset F^n$ and it suffices to confirm that $\operatorname{OutLim} F^n \subset F$. Take $f \in \operatorname{OutLim} F^n$. There exists $f^k \in F^{n_k} \to f$. Since $\int x f^k(x)dx \in C^{n_k}$, that integral converges to $\int x f(x)dx$ by Proposition 4.8, and the set-convergence C^n to C allow us to conclude that $\int x f(x)dx \in C$. □

Proof of Theorem 5.1. Let $\varphi, \varphi^\nu : F^0 \to (-\infty, \infty]$ be given by $\varphi(f) = n^{-1}\sum_{j=1}^n \psi(x^j, f) + \pi(f)$ if $f \in F$ and $\varphi(f) = \infty$ otherwise; and $\varphi^\nu(f) = n^{-1}\sum_{j=1}^n \psi(x^j, f) + \pi^\nu(f)$ if $f \in F^\nu$ and

$\varphi^\nu(f) = \infty$ otherwise. We start by showing that φ^ν epi-converges to φ. Let $f^\nu \in F^0 \to f$. If $f \in F$, then

$$\operatorname{liminf} \varphi^\nu(f^\nu) \geq \frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f) + \pi(f) = \varphi(f).$$

If $f \notin F$, then because F is closed we must have that $f^\nu \notin F^\nu$ for sufficiently large ν. Thus, $\operatorname{liminf} \varphi^\nu(f^\nu) = \varphi(f) = \infty$. Next, let $f \in F$. There exists $f^\nu \in F^\nu \to f$ because F^ν set-converges to F. Then,

$$\operatorname{limsup} \varphi^\nu(f^\nu) = \operatorname{limsup} \Big(\frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f^\nu) + \pi^\nu(f^\nu) \Big) = \varphi(f).$$

This is sufficient for φ^ν epi-converging to φ. Reasoning along the lines of those in the proof of Theorem 3.10 yields (i).

For (ii), we recognize that the additional condition on F^0 ensures that it is compact. Thus, $\{f^\nu, \nu \in I\!N\}$ in the statement of the theorem must have a cluster point. Every such cluster point must be in $\varepsilon^\infty\text{-}\operatorname{argmin}_{f \in F} \varphi(f)$. Let $\delta = \varepsilon - \varepsilon^\infty$, which is positive. Since F^0 is compact, π^ν converges uniformly to π. Hence, there exists $\bar{\nu} \in I\!N$ such that $\pi(f^\nu) \leq \pi^\nu(f^\nu) + \delta/3$, $\varepsilon^\nu \leq \varepsilon^\infty + \delta/3$, and, in view of epi-convergence, $\inf_{f \in F^\nu} \varphi^\nu(f) \leq \inf_{f \in F} \varphi(f) + \delta/3$ for all $\nu \geq \bar{\nu}$. Since $F^\nu \subset F$, we then have

$$\varphi(f^{\bar{\nu}}) = \frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f^{\bar{\nu}}) + \pi(f^{\bar{\nu}}) \leq \frac{1}{n} \sum_{j=1}^{n} \psi(x^j, f^{\bar{\nu}}) + \pi^{\bar{\nu}}(f^{\bar{\nu}}) + \delta/3$$
$$\leq \inf_{f \in F^{\bar{\nu}}} \varphi^{\bar{\nu}}(f) + \varepsilon^{\bar{\nu}} + \delta/3 \leq \inf_{f \in F} \varphi(f) + \varepsilon^\infty + \delta = \inf_{f \in F} \varphi(f) + \varepsilon,$$

which establishes the claim. □