DERIVING THE CONTINUITY OF MAXIMUM-ENTROPY BASIS FUNCTIONS VIA VARIATIONAL ANALYSIS*

N. SUKUMAR† AND R. J-B WETS‡

Abstract. In this paper, we prove the continuity of maximum-entropy basis functions using Variational Analysis techniques. The use of information-theoretic variational principles to derive basis functions is a recent development. In this setting, data approximation is viewed as an inductive inference problem, with the basis functions being synonymous with a discrete probability distribution and the polynomial reproducing conditions acting as the linear constraints. For a set of distinct nodes $\{x^i\}_{i=1}^n$ in $\mathbb{R}^d$, the convex approximation of a function $u(x)$ is: $u^h(x) = \sum_{i=1}^n p_i(x)u_i$, where $\{p_i\}_{i=1}^n$ are non-negative basis functions, and $u^h(x)$ must reproduce affine functions: $\sum_{i=1}^n p_i(x) = 1$, $\sum_{i=1}^n p_i(x)x^i = x$. Given these constraints, we compute $p_i(x)$ by minimizing the relative entropy functional (Kullback-Leibler distance), $D(p\|m) = \sum_{i=1}^n p_i(x)\ln\big(p_i(x)/m_i(x)\big)$, where $m_i(x)$ is a known prior weight function distribution. To prove the continuity of the basis functions, we appeal to the theory of epi-convergence.

Key words. maximum entropy, relative entropy, convex approximation, meshfree methods, epi-convergence

AMS subject classifications. 65N30, 65K10, 90C25, 62B10, 26B25

1. Background and formulation.

Consider a set of distinct nodes in $\mathbb{R}^d$ that are located at x^i $(i = 1, 2, \ldots, n)$, with $D = \text{con}(x^1, \ldots, x^n) \subset \mathbb{R}^d$ denoting the convex hull of the nodal set (Figure 1.1). For a real-valued function $u(x) : D \to \mathbb{R}$, the numerical approximation for $u(x)$ is written as

$$u^h(x) = \sum_{i=1}^n p_i(x)u_i, \tag{1.1}$$

where $p_i(x)$ is the basis function associated with node i, and u_i are coefficients. If $p_i(x)$ is a cardinal basis, $p_i(x^j) = \delta_{ij}$, then $u^h(x^i) = u(x^i) = u_i$.

In the univariate case, Lagrange and spline bases are well-known, whereas for multivariate approximation, tensor-product splines, moving least squares (MLS) approximates [17] and radial basis functions [30] are popular. The need for scattered data approximation arises in many fields, for example, curve and surface fitting, computer graphics and geometric modeling, finite elements, and meshfree methods. Over the past decade, meshfree approximation schemes have been adopted in Rayleigh-Ritz (Galerkin) methods for the modeling and simulation of physical phenomena; see [4] for a review of meshfree methods and [28] for a review of meshfree basis functions. For second-order partial differential equations (PDEs), approximates that possess constant and linear precision are sufficient for convergence in a Galerkin method, cf., for example, [25, Chapter 2]:

$$\forall x, \quad \sum_{i=1}^n p_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^n p_i(x)x^i = x. \tag{1.2}$$

*Research supported in part by the National Science Foundation through grants CMMI-0626481 and DMS-0205699.

†Department of Civil and Environmental Engineering, University of California, Davis, CA 95616 (nsukumar@ucdavis.edu).

‡Department of Mathematics, University of California, Davis, CA 95616 (rjbwets@ucdavis.edu).

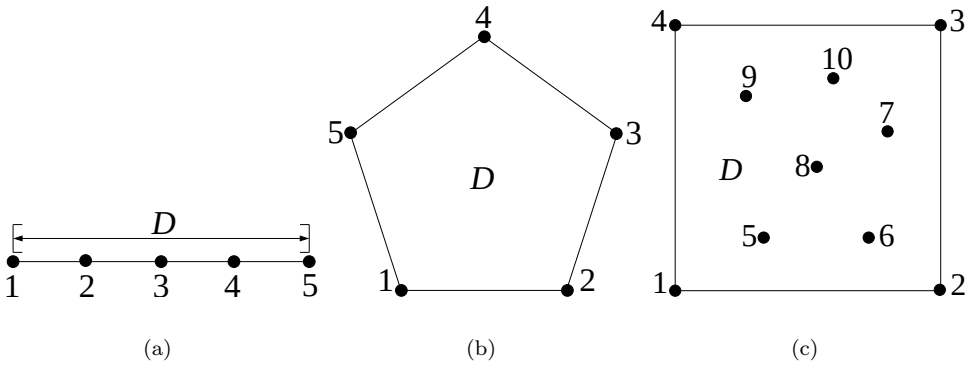


FIG. 1.1. *Nodal locations x^i. (a) One-dimension; (b) Pentagon; and (c) Scattered nodes within a square.*

Furthermore, if the non-negative restriction is imposed on the basis functions (convex combination), namely

$$p_i(x) \geq 0 \quad \forall i,\, x, \tag{1.3}$$

then (1.1) is a convex approximation scheme [1] with many desirable properties—satisfies the convex hull property, not prone to the Runge phenomena, interior nodal basis functions $p_i(x)$ ($x^i \notin \text{bdry}\, D$) vanish on bdry D, which facilitates the imposition of linear Dirichlet boundary conditions in a Galerkin method, and in addition optimal conditioning can be established for non-negative basis functions [8, 19].

In meshfree Galerkin methods, an approximation of the form in (1.1) is used, with moving least squares being the most common choice. A recent development in this direction has been the construction of maximum-entropy (`MAXENT`) approximates [1, 26, 27]; continuity was obtained by Arroyo and Ortiz [1] when the prior distributions are Gaussian. In this paper, we rely on *Variational Analysis* techniques, in particular on the theory of *epi-convergence*, to establish the continuity of maximum-entropy basis functions for *any* continuous prior distribution.

1.1. Minimum relative entropy principle. In information theory [7], the notion of entropy as a measure of uncertainty or incomplete knowledge was introduced by Shannon [22]. The Shannon entropy of a discrete probability distribution is:

$$H(p) = <-\ln p> = -\sum_{i=1}^{n} p_i \ln p_i, \tag{1.4}$$

where $<\cdot>$ is the expectation operator, $p_i \equiv p(x^i)$ is the probability of the occurrence of the event x^i, $p \ln p \doteq 0$ if $p = 0$, and the above form of H satisfies the axiomatic requirements of an uncertainty measure, cf., for example, [14, Chapter 1].

Jaynes used the Shannon entropy measure to propose the principle of maximum entropy [11], in which it was shown that maximizing entropy provides the least-biased statistical inference when insufficient information is available. It was later recognized that for H to be invariant under invertible mappings of x, the general form of the

entropy should be [12, 15, 23]

$$H(p,m) = -\int p(x)\ln\left(\frac{p(x)}{m(x)}\right)dx, \quad \text{or} \quad H(p,m) = -\sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right), \tag{1.5}$$

where m is a prior distribution that plays the role of a p-estimate. In the literature, the quantity $D(p\|m) = -H(p,m)$ is also referred to as the Kullback-Leibler distance (directed- or I-divergence) [16], and the variational principle is known as the principle of minimum relative entropy [23]. If a uniform prior, $m_i = 1/n$, is used in (1.5), then the Shannon entropy (modulo a constant) given in (1.4) is recovered. The non-negativity of the relative entropy, $D(p\|m) \geq 0$, is readily derived from Jensen's inequality (cf., for example, [7, p.25]),

Given a set of $\ell + 1$ linear constraints on an unknown probability distribution p and a prior m, which is an estimate for p, the minimum relative entropy principle is a rule for the most consistent (minimum-distance or -discrepancy from the prior m) assignment of the probabilities p_i [12]:

$$\min_{p\in\boldsymbol{R}^n_+}\left(D(p\|m) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{m_i}\right)\right) \quad \text{so that } \sum_{i=1}^{n} p_i = 1, \tag{1.6a}$$

$$\sum_{i=1}^{n} p_i g_r(x^i) = <g_r(x)>, \quad r = 1, 2, \ldots, \ell, \tag{1.6b}$$

where $g_r(x)$ and $<g_r(x)>$ are known, and $\mathbb{R}^n_+$ is the non-negative orthant.

The initial emphasis of the principle of maximum entropy was in equilibrium and non-equilibrium statistical mechanics [12], but it is equally applicable to any problem in inductive inference. The interested reader can refer to [13] and [24] for the Bayesian perspective on probability theory and rationale inference. The maximum entropy and minimum relative entropy principles have found applications in many areas of science and engineering—image reconstruction [10], natural language modeling [5], microstructure reconstruction [18], and non-parametric supervised learning [9] are a few examples.

Variational principles, which are used in finite element formulations, conjugate gradient methods, graphical models, dynamic programming, and statistical mechanics, also have strong roots in data approximation. For instance, kriging, thin-plate splines, B-splines, radial basis functions [30], MLS approximates [17], and Delaunay interpolates [20] are based on the extremum of a functional. In the same spirit, we now present the variational formulation to construct entropy approximates, and in so doing, demonstrate its potential merits as a basis for the solution of PDEs.

1.2. Variational formulation for entropy approximates. To obtain the maximum-entropy principle, the Shannon entropy functional and a modified entropy functional, were used in [26] and [1], respectively. In [27], as a unifying framework and generalization, the relative entropy functional with a prior was used—a uniform prior leads to Jaynes's maximum-entropy principle and use of a Gaussian (radial basis function) prior, $m_i(x) = \exp(-\beta|x^i - x|^2)$, results in the entropy functional considered in [1]. The prior in the present context is a nodal weight function, and the variational principle in effect provides a 'correction' that minimally modifies the weight functions to form basis functions that also satisfy the linear constraints. Clearly, if $m_i(x)$ *a priori* satisfies all the constraints, then one obtains $p_i(x) = m_i(x)$ for all i.

The flexibility of choosing different prior distributions (for example, radial basis functions, compactly-supported weight functions used in MLS, etc.) within the minimum relative entropy formalism would lead to the construction of a wider-class of convex approximation schemes. The parallels between the conditions on p_i in (1.2) and (1.3) and those on p_i in a `MAXENT` formulation are evident. Unlike univariate Bernstein basis functions (terms in the binomial expansion), where a probabilistic interpretation in relation to the binomial distribution [24, Chapter 5] is natural, here the connection is less transparent. Referring to the nodal sets shown in Figure 1.1, the basis function value $p_i(x)$ is viewed as the 'probability of influence of a node i at x.' With a uniform prior, global basis functions are obtained, which do not lead to sparse system matrices in the numerical solution of PDEs. With a compactly-supported prior, the basis functions $p_i(x)$ also inherit the support properties of the prior and hence are suitable in the Galerkin solution for PDEs. Entropic regularization with a prior is a novel approach to construct convex approximation schemes with many desirable properties.

The variational formulation for entropy approximates is: find $x \mapsto p(x) : \mathbb{R}^d \to \mathbb{R}^n_+$ as the solution of the following constrained convex optimization problem:

$$\min_{p \in \mathbb{R}^n_+} f(x;p), \quad f(x;p) = \sum_{i=1}^{n} p_i(x) \ln\left(\frac{p_i(x)}{m_i(x)}\right), \tag{1.7a}$$

subject to the constraint set from (1.2) and (1.3):

$$\kappa(x) = \Big\{ p \in \mathbb{R}^n_+ \;\Big|\; \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i x^i = x \Big\}, \tag{1.7b}$$

where $m_i(x)$ is a prior estimate, and the constraints form an under-determined linear system. By introducing the Lagrange multipliers, one can write the solution of the variational problem as:

$$p_i(x) = \frac{Z_i(x)}{Z(x)}, \quad Z_i(x) = m_i(x) \exp(-x^i \cdot \lambda),$$

where $\lambda \in \mathbb{R}^d$, and $Z(x) = \sum_j Z_j(x)$ is known as the partition function in statistical mechanics. The $p_i(x)$ in the preceding equation must satisfy the d linear constraints in (1.7b). This yields d nonlinear equations. On using shifted nodal coordinates $\tilde{x}^i = x^i - x$ and considering the dual formulation, we can write the solution for the Lagrange multipliers as (cf., for example, [21, Exercise 11.12] and [6, p.222])

$$\lambda = \arg\min \ln Z(\lambda^t),$$

where Z is appropriately redefined. Convex optimization algorithms (gradient descent, Newton's method) are suitable to compute these basis functions. Numerical experimentation suggest that such basis functions may very well be continuous on D [1, 26], and this will be confirmed here, by relying on Variational Analysis techniques.

2. Continuity of the basis functions. One can always represent an optimization problem, involving constraints or not, as one of minimizing an extended real-valued function. In the case of a constrained-minimization problem, simply redefine the effective objective as taking on the value ∞ outside the feasible region, the set determined by the constraints. In this framework, the canonical problem can

be formulated as one of minimizing on all of $\mathbb{R}^n$ an extended real-valued function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. Approximation issues can consequently be studied in terms of the convergence of such functions. This has lead to the notion of *epi-convergence*, cf. [2, 3] and [21, Chapter 7]; the latter will serve here as our basic reference. We provide a very brief survey and some relevant refinements of this theory.

Thus, at a conceptual level, it's convenient to think of optimization problems as elements of

$$\operatorname{fcns}(\mathbb{R}^n) = \left\{ f\colon \mathbb{R}^n \to \overline{\mathbb{R}} \right\}$$

the set of extended real-valued functions that are defined on *all* of $\mathbb{R}^n$, even allowing for the possibility that they are nowhere finite-valued; definitions, properties, limits, etc., usually do not refer specifically to the domain on which they are finite. The *effective domain* of f is $\operatorname{dom} f = \left\{ x \in \mathbb{R}^n \,\middle|\, f(x) < \infty \right\}$. The *epigraph* of a function f is the set of all points in $\mathbb{R}^{n+1}$ that lie on or above the graph of f, $\operatorname{epi} f = \left\{ (x,\alpha) \in \mathbb{R}^{n+1} \,\middle|\, \alpha \geq f(x) \right\}$. A function f is *lsc* (= *lower semicontinuous*) if and only if its epigraph is closed as a subset of $\mathbb{R}^{n+1}$, i.e., $\operatorname{epi} f = \operatorname{cl}(\operatorname{epi} f)$ with cl denoting closure [21, Theorem 1.6]. The lsc-regularization of f is $\operatorname{cl} f$ defined by the identity $\operatorname{epi} \operatorname{cl} f = \operatorname{cl} \operatorname{epi} f$.

Definition 2.1. (epi- and tight epi-convergence). *Let* $\left\{ f, f^\nu,\, \nu \in \mathbb{N} \right\}$ *be a collection of functions in* $\operatorname{fcns}(\mathbb{R}^n)$. *Then,* $f^\nu \xrightarrow{e} f$ *if and only if the following conditions are satisfied:*

(a) $\forall\, x^\nu \to x$, $\operatorname{liminf}_\nu f^\nu(x^\nu) \geq f(x)$,

(b) $\forall x$, $\exists\, x^\nu \to x$ *such that* $\operatorname{limsup}_\nu f^\nu(x^\nu) \leq f(x)$.

The sequence epi-converges tightly *to* f *if, in addition, for all* $\epsilon > 0$, *there exist a compact set* B_ϵ *and an index* ν_ϵ *such that*

$$\forall\, \nu \geq \nu_\epsilon : \quad \inf\nolimits_{B_\epsilon} f^\nu \leq \inf f^\nu + \epsilon.$$

Note that functions can be 'epi-close' while 'pointwise-far' (measured, for example, in term of the ℓ^∞-norm); e.g., consider the two step-functions $f(x) = 0$ if $x < 0$, $f(x) = 1$ when $x \geq 0$ and $g(x) = f(x - \epsilon)$ with $\epsilon > 0$ arbitrarily small.

The name 'epi-convergence' is attached to this convergence notion because it coincides ([21, Proposition 7.2]) with the *set-convergence*, in the Painlevé-Kuratowski sense [21, §4.B] of the epigraphs. It's known that (i) whenever C is a limit-set, it's *closed* [21, Proposition 4.4], (ii) $C = \emptyset$ if and only if the sequence C^ν eventually 'escapes' from any bounded set [21, Corollary 4.11], and (iii) if the sequence $C^\nu \to C$ consists of convex sets, then also C is convex [21, Proposition 4.15]. This means that when $f^\nu \xrightarrow{e} f$, (i) f is lsc, (ii) $f \equiv \infty (\operatorname{dom} f = \emptyset)$ if and only if given any $\kappa > 0$, $f^\nu \geq \kappa$ for ν large enough, and (iii) the epi-limit of convex functions is convex, if it exists.

Theorem 2.2. (convergence of the minimizers and infimums). *Let* $f^\nu \xrightarrow{e} f$, *all in* $\operatorname{fcns}(\mathbb{R}^n)$, *with* $\inf f$ *finite. If* $f^\nu \xrightarrow{e} f$, $x^k \in \operatorname{argmin} f^{\nu_k}$ *for some subsequence* $\{\nu_k\}_{k\in\mathbb{N}}$ *and* $x^k \to \bar{x}$, *then* $\bar{x} \in \operatorname{argmin} f$ *and* $\min f^{\nu_k} \to \min f$.[1]

If $\operatorname{argmin} f$ *is a singleton, then every convergent subsequence of minimizers converges to* $\operatorname{argmin} f$.

They epi-converge tightly if and only if $\inf f^\nu \to \inf f$.

[1] one writes min when the infimum is actually attained

Proof. The two first assertions follow from [21, Proposition 7.30, Theorem 7.33] and one can deduce the last one from [21, Theorem 7.31]. □

Let's conclude this review by a compilation of the facts that are going to be of immediate relevance to the problem at hand.

COROLLARY 2.3. (epi-convergence under strict convexity). *Suppose* $\{f^\nu : \mathbb{R}^n \to (-\infty, \infty]\}_{\nu \in \mathbb{N}}$ *be a collection of convex functions such that*

(a) *for all* ν, $\operatorname{dom} f^\nu \subset B$ *where* B *and each* $\operatorname{dom} f^\nu$ *are compact,*

(b) *the functions* f^ν *are finite-valued, lsc and strictly convex on* $\operatorname{dom} f^\nu$. *Then, for all* ν, $\emptyset \neq \operatorname{argmin} f^\nu$ *is a singleton,*

Moreover, if $f^\nu \xrightarrow{e} f$ *and* $\operatorname{argmin} f$ *is also a singleton, then* $\operatorname{argmin} f^\nu \to \operatorname{argmin} f$.

Proof. In view of (a) and (b), for each ν the minimization of f^ν is equivalent to minimizing a finite-valued, lsc, strictly convex function on a compact set, and such a problem always has a unique solution. Moreover, because for all ν, $\operatorname{dom} f^\nu$ is a (compact) subset of the compact set B, $f^\nu \xrightarrow{e} f$ implies that they epi-converge tightly. The convergence of $\operatorname{argmin} f^\nu \to \operatorname{argmin} f$ follows from combining the two last assertions of Theorem 2.2. □

Our task, now, is to show that the continuity of the basis functions can be derived as a consequence of this corollary. We begin with the strict convexity of the criterion function. The Kullback-Leibler criterion is a separable function, i.e.,

$$k(x;p) = \sum_{i=1}^{n} k_i(x;p_i) \text{ where } k_i(x;p_i) = p_i \ln(p_i/m_i(x)),$$

and its properties can be directly derived from those of the 1-dimensional functions $k_i(x;\cdot) : \mathbb{R}_+ \to [0,\infty]$.

- When $m_i(x) > 0$, $k_i(x,\cdot)$ is finite-valued, continuous and strictly convex on $\mathbb{R}_+$; recall that $0\ln(0) = 0$. Indeed, the second derivative on $(0,\infty)$ is $1/p_i > 0$ that implies strict convexity [21, Theorem 2.13(c)]. The quantity $p_i \ln(p_i/m_i(x))$ is strictly increasing and converges to 0 as $p_i \searrow 0$ yielding both strict convexity and continuity on $\mathbb{R}_+$.
- When $m_i(x) = 0$, $k_i(x;p_i) = \infty$ unless $p_i = 0$ and then $k_i(x;0) = 0$.

It's conceivable, but certainly not reasonable, that the (continuous) weight functions $\{m_i : \mathbb{R}^n \to \mathbb{R}_+,\ i = 1,\dots,n\}$ have been chosen so that for some $x \in D$, $m_i(x) = 0$ for all $i = 1,\dots,n$. In such a situation, in the process of minimizing the Kullback-Leibler criterion, we would be lead to choose $p = 0$ and, of course, this would make it impossible to satisfy the constraint $\sum_{i=1}^n p_i = 1$, i.e., the problem, so formulated, would be infeasible! This brings us to the following assumption: Let

- $\text{s-supp}\, m_i = \{x \in \mathbb{R}^d \mid m_i(x) > 0\}$ denote the *strict support* of m_i, and
- $\operatorname{supp} m_i = \operatorname{cl}(\text{s-supp}\, m_i)$ the *support* of m_i.

ASSUMPTION 2.1. (well-posed assumption). *For each* $i = 1,\dots,n$, *the function* $m_i : \mathbb{R}^n \to \mathbb{R}_+$ *is continuous such that* $\text{s-supp}\, m_i$, *and consequently also* $\operatorname{supp} m_i$, *is non-empty.*[2] *Moreover, with* $I_{=0} = \{i \mid m_i(x) = 0\}$ *and* $I_{>0} = \{i \mid m_i(x) > 0\}$,

$$\text{for all } x \in D: \quad x \in \operatorname{con}(x^i \mid i \in I_{>0}).$$

[2] Note that the continuity of m_i implies that $\text{s-supp}\, m_i$ is an open subset of $\mathbb{R}^d$, and so is $\bigcup_{i=1}^n \text{s-supp}\, m_i$.

This assumption requires that every $x \in D$ can be obtained as a convex combination of some subcollection of the nodal locations x^i that are associated with weight functions m_i that have $m_i(x) > 0$. In particular, this implies that $\kappa(x)$ is never empty, or equivalently, that the constraints (7b) are certainly satisfied whenever $x \in D$.

PROPOSITION 2.4. (the Kullback-Leibler (KL) criterion). *Under the well-posed Assumption 2.1, for all $x \in D$, the KL-criterion $p \mapsto k(x;p) = \sum_{i=1}^n p_i \ln(p_i/m_i(x))$ is a strictly convex, lsc function on $\mathbb{R}^n_+$, taking into account the identity $0\ln(0) = 0$.*

Proof. Convexity is well-known, see [7, p.30], [21, Exercise 3.51], for example. Again, with $I_{=0} = \{i \,|\, m_i(x) = 0\}$ and $I_{>0} = \{i \,|\, m_i(x) > 0\}$,

$$k(x;p) = \sum_{i \in I_{=0}} k_i(x;p_i) + \sum_{i \in I_{>0}} k_i(x;p_i),$$

$\operatorname{dom} k(x;\cdot) = \prod_{I_{=0}}\{0\} \times \prod_{I_{>0}} \mathbb{R}_+$ and $I_{>0}$ non-empty by Assumption 2.1. From our analysis of the functions $k_i(x;\cdot)$, it follows that $k(x;\cdot)$ is strictly convex, continuous on its effective domain $\operatorname{dom} k(x;\cdot)$. □

The tools are now at hand to derive our main result.

THEOREM 2.5. (continuity of the basis functions). *For $x \in D$, as in the formulation of* ***MAXENT*** *(1.7), let*

$$f(x;p)) = \begin{cases} \sum_{i=1}^n p_i \ln(p_i/m_i(x)) & \text{if } p \in \kappa(x), \\ \infty & \text{otherwise,} \end{cases}$$

where,

$$\kappa(x) = \{p \in \mathbb{R}^n_+ \,|\, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i x^i = x\},$$

and

$$p(x) = (p_1(x), \ldots, p_n(x)) = \operatorname{argmin} f(x;\cdot).$$

Under the well-posed Assumption 2.1, when $x^\nu \to \bar{x}$ with $x^\nu \in D$, $\kappa(\bar{x})$ is non-empty and

$$f(x^\nu;\cdot) \xrightarrow{e} f(\bar{x};\cdot) \quad \text{and} \quad p(x^\nu) \to p(\bar{x}).$$

In other words, the basis functions $p(\cdot)$ are continuous on D.

Proof. Since for all $x \in D$, $\kappa(x)$ is a compact, non-empty subset of the unit simplex $\Delta = \{p \in \mathbb{R}^n_+ \,|\, \sum_{i=1}^n p_i = 1\}$, it follows that for all $x \in D$, $\operatorname{dom} f(x;\cdot) \subset \Delta$ and, consequently, condition (a) of Corollary 2.3 is trivially satisfied. The rest of the proof is concerned with condition (b) and the epi-convergence of the sequence $f(x^\nu;\cdot)$ to $f(\bar{x};\cdot)$ when $x^\nu \to \bar{x}$.

The functions $f(x^\nu;\cdot)$ and $f(\bar{x};\cdot)$ can be written as $k(x^\nu;\cdot) + \iota_{\kappa(x^\nu)}$ and $k(\bar{x};\cdot) + \iota_{\kappa(\bar{x})}$ where $k(x;p)$ is the Kullback-Leibler criterion defined on $\mathbb{R}^n_+$ and ι_C is the indicator function of the set $C \subset \mathbb{R}^n$ with $\iota_C = 0$ on C and otherwise, $\iota_C = \infty$ on $\mathbb{R}^n \setminus C$.

The epi-convergence of $f(x^\nu;\cdot)$ to $f(\bar{x};\cdot)$ follows from [21, Theorem 7.46(b)] which asserts that the sum of two sequences of functions epi-converge to the sum of their limits, if one sequence epi-convergences and the other converges continuously.

To obtain the epi-convergence of the indicator functions, or equivalently ([21, Proposition 7.4(f)]), the set convergence of the sets $\kappa(x^\nu) \to \kappa(\bar{x})$ with $\kappa(\bar{x}) \neq \emptyset$, we exploit the fact that these are polyhedral sets and that, on the bounded polyhedral set $D = \mathrm{con}(x^1, \ldots, x^n) \subset I\!R^d$, the mapping $x \mapsto \kappa(x)$ is Lipschitz continuous with respect to the Pompeiu-Hausdorff distance $d\!l_\infty$, i.e.,

$$\forall\, x, x' \in D\colon \quad d\!l_\infty(\kappa(x), \kappa(x')) \leq M|x - x'|$$

for some constant $M > 0$; here $|\cdot|$ denotes the Euclidean norm, cf. [29, Theorem 1], also [21, Example 9.35]. Of course, this means that κ is continuous on D and, in particular, for any $x^\nu \to \bar{x}$ in D, given any sequence $p^\nu \in \kappa(x^\nu) \to \bar{p}$, then $\bar{p} \in \kappa(\bar{x})$.

Thus, to assert continuous convergence of the functions $k(x^\nu;\cdot)$ to $k(\bar{x})$, one needs to show that for such pairs (x^ν, p^ν): $k(x^\nu;p^\nu) \to k(\bar{x};\bar{p})$. Let $I_{=0} = \big\{i \,\big|\, m_i(\bar{x}) = 0\big\}$ and $I_{>0} = \big\{i \,\big|\, m_i(\bar{x}) > 0\big\}$. By Assumption 2.1, $\kappa(\bar{x}) \bigcap (\bigcup_{I_{>0}} \text{s-supp}\, m_i) \neq \emptyset$. Furthermore, the open set $\bigcup_{I_{>0}} \text{s-supp}\, m_i$ not only includes $\bar{x}$ but also x^ν for all ν large enough. Thus, for all $i \in I_{>0}$, $p_i^\nu \ln(p_i^\nu)/m_i(x^\nu)) \to \bar{p}_i \ln(\bar{p}_i)/m_i(\bar{x})$. When $i \in I_{=0}$, again for ν large enough, the $p_i^\nu = 0 = \bar{p}_i$, otherwise the corresponding vectors p^ν and $\bar{p}$ would not belong to $\mathrm{dom}\, k(x^\nu;\cdot)$ or $\mathrm{dom}\, k(\bar{x};\cdot)$. Hence, $k(x^\nu;p^\nu) \to k(\bar{x};\bar{p})$. So, $f(x^\nu;\cdot) \to f(\bar{x},\cdot)$.

There only remains to observe that, for ν large enough, $\mathrm{argmin}\, f(x^\nu;\cdot)$ is unique, i.e., for $i \notin I_{>0}$, $p_i^\nu(x^\nu) = 0$, whereas for $i \in I_{>0}$, $p_i^\nu(x^\nu) = \mathrm{argmin}_{p_i \geq 0} p_i \ln(p_i/m_i(x^\nu))$; the strict convexity guarantees that argmin is a singleton. Since the same holds for $\bar{x}$, we are in the framework of Corollary 2.3, and thus $p(x^\nu) = \mathrm{argmin}\, f(x^\nu;\cdot) \to \mathrm{argmin}\, f(\bar{x};\cdot) = p(\bar{x})$. □

3. Numerical experiments. To illustrate Theorem 2.5, we present basis functions plots to confirm the continuity of maximum-entropy basis functions. First, 1-dimensional basis functions are considered, and then 2-dimensional basis function plots are presented.

To demonstrate a simple closed-form computation, consider 1-dimensional approximation in $D = [0, 1]$ with three nodes located at $x_1 = 0$, $x_2 = 1/2$, and $x_3 = 1$. On using (1.7), the solution for $p_i(x)$ is obtained by solving a quadratic equation:

$$p_1(x) = \frac{1}{Z}, \quad p_2(x) = \frac{\eta}{Z}, \quad p_3(x) = \frac{\eta^2}{Z}, \quad \eta \equiv \eta(x) = \frac{2x - 1 + \sqrt{12x(1-x)+1}}{4(1-x)},$$

where $Z = 1 + \eta + \eta^2$. These basis functions are presented in Figure 3.1a. For four equi-spaced nodes in $[0, 1]$, a cubic equation must be solved. In general, a numerical method is required to compute these basis functions; in our computations, we use a 1-dimensional `Matlab`$^{\mathrm{TM}}$ implementation, whereas in two dimensions, a gradient descent algorithm [26, p.2165] is adopted. In Figure 3.1, basis function plots on a uniform grid consisting of three nodes and five nodes (nodal locations are shown in Figure 1.1a) are depicted. The plots are presented for a Gaussian prior distribution, $m_i(x) = \exp(-\beta(|x^i - x|^2)$, with varying β. The value $\beta = 0$ corresponds to a uniform prior, and for large β (theoretically when $\beta \to \infty$), the entropy basis functions tend to the finite element Delaunay interpolant [1]. From Figures 3.1a and 3.1d, we observe that nodal interpolation is realized on the boundary but not at the interior nodes. However, as β is increased the support of the basis functions shrinks and the basis functions become closer to being an interpolant at the interior nodes. For $\beta = 100$, the entropy basis functions are proximal to piece-wise linear finite element basis functions

(Figures 3.1c and 3.1f). The plots in Figure 3.1 evince the continuity of the basis functions, which provides numerical evidence in support of the theoretical proof in Theorem 2.5.

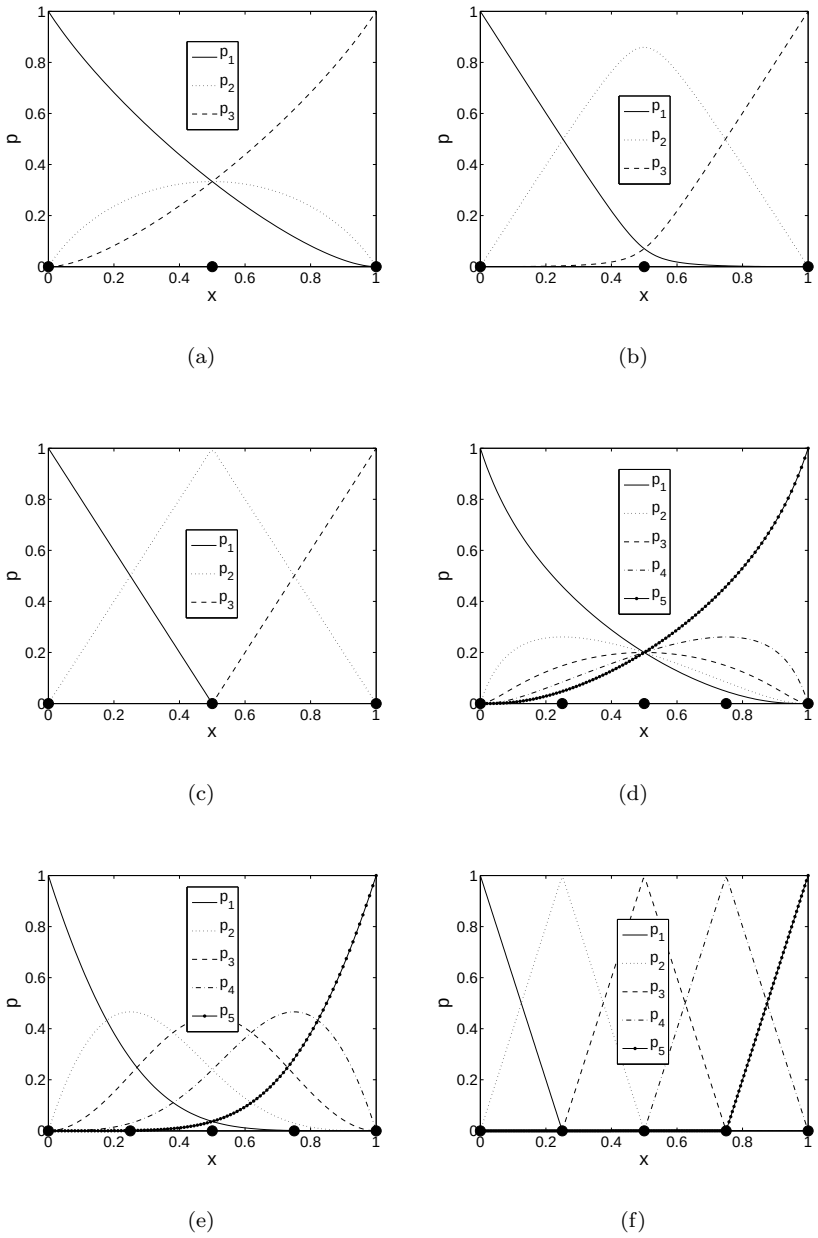


FIG. 3.1. *Entropy basis functions with a Gaussian prior. (a)–(c)* $n = 3$ *and* $\beta = 0$, 10, 100*; and (d)–(f)* $n = 5$ *and* $\beta = 0$, 10, 100. *The nodal locations along the x-axis are depicted by filled circles.*

In Figure 3.2a, a contour plot of $p_1(x)$ for node 1 in a regular pentagon (see

Figure 1.1b for the nodal locations) is shown, whereas in Figure 3.2b, the 3D plot is illustrated. The variation of the maximum entropy within the pentagon is depicted in Figure 3.2c, with the maximum value of $\ln 5$ being attained at the centroid of the pentagon. The basis function $p_1(x)$ satisfies the cardinal property, $p_i(x^j) = \delta_{ij}$, which is also met by all n nodal basis functions in a convex polygon [26]. Next, we consider the grid shown in Figure 1.1d, where $D = [0,1]^2$. The basis function for nodes 1 and 8 are plotted using a uniform prior, a Gaussian prior with $\beta = 20$, and a compactly-supported C^2 quartic radial basis function as a prior. The quartic prior is given by $m_i(r) = 1 - 6r^2 + 8r^3 - 3r^4$ if $r = |x^i - x| \leq 1$, and zero otherwise. The contour plots are illustrated in Figure 3.3, and once again we observe that the basis functions are continuous in D. Furthermore, the interior basis functions (for example, $p_8(x)$) vanish on bdry D, which enables the direct imposition of Dirichlet boundary conditions in Galerkin methods [1]. The 1- and 2-dimensional basis function plots provide numerical proof in support of Theorem 2.5, thereby establishing the continuity of $p_i(x)$ for $x \in D$.

REFERENCES

[1] M. Arroyo and M. Ortiz, *Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods*, International Journal for Numerical Methods in Engineering, 65 (2006), pp. 2167–2202.

[2] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable Mathematics Series, Pitman, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, 1990.

[4] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl, *Meshless methods: An overview and recent developments*, Computer Methods in Applied Mechanics and Engineering, 139 (1996), pp. 3–47.

[5] A. L. Berger, S. A. DellaPietra, and V. J. DellaPietra, *A maximum entropy approach to natural language processing*, Computational Linguistics, 22 (1996), pp. 39–71.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, NY, 1991.

[8] R. T. Farouki and T. N. T. Goodman, *On the optimal stability of the Bernstein basis*, Mathematics of Computation, 65 (1996), pp. 1553–1566.

[9] M. R. Gupta, *An Information Theory Approach to Supervised Learning*, ph.D. thesis, Department of Electrical Engineering, Stanford University, Palo Alto, CA, U.S.A., March 2003.

[10] Skilling J. and R. K. Bryan, *Maximum entropy image reconstruction: general algorithm*, Monthly Notices of the Royal Astronomical Society, 211 (1984), pp. 111–118.

[11] E. T. Jaynes, *Information theory and statistical mechanics*, Physical Review, 106 (1957), pp. 620–630.

[12] ———, *Information theory and statistical mechanics*, in Statistical Physics: The 1962 Brandeis Lectures, K. Ford, ed., New York, 1963, W. A. Benjamin, pp. 181–218.

[13] ———, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[14] A. Khinchin, *Mathematical Foundations of Information Theory*, Dover, New York, N.Y., 1957.

[15] S. Kullback, *Information Theory and Statistics*, Wiley, New York, NY, 1959.

[16] S. Kullback and R. A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics, 22 (1951), pp. 79–86.

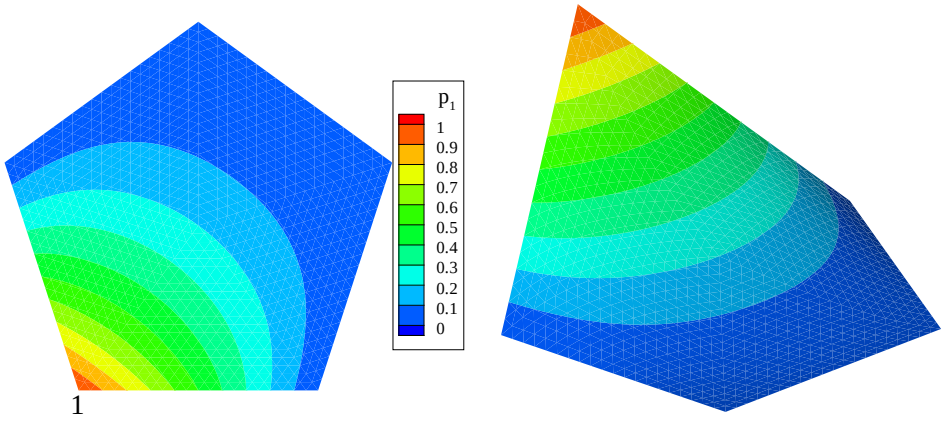
[17] P. Lancaster and K. Salkauskas, *Surfaces generated by moving least squares methods*, Mathematics of Computation, 37 (1981), pp. 141–158.

[18] R. W. Minich, C. A. Schuh, and M. Kumar, *Role of topological constraints on the statistical properties of grain boundary networks*, Physical Review B, 66 (2003), p. 052101.

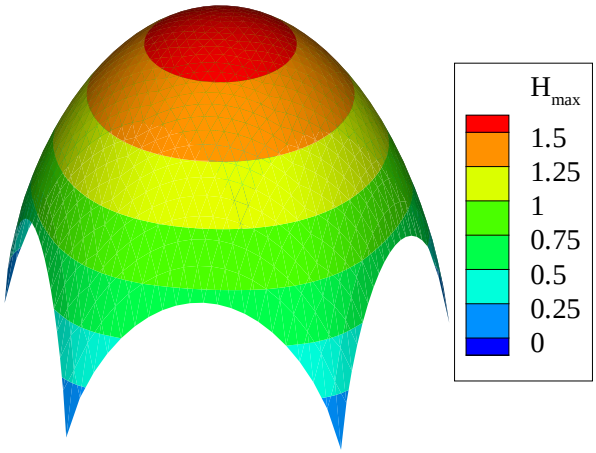
[19] J. M. Peña, *B-splines and optimal stability*, Mathematics of Computation, 66 (1997), pp. 1555–1560.

[20] V. T. Rajan, *Optimality by the Delaunay triangulation in R^d*, Discrete and Computational Geometry, 12 (1994), pp. 189–202.

[21] R. T. Rockafellar and R. J-B Wets, *Variational Analysis*, Springer-Verlag, Heidelberg,



(a) (b)



(c)

FIG. 3.2. *Entropy basis function $p_1(x)$ and variation of maximum entropy within a regular pentagon. (a) Contour plot; (b) 3D plot; and (c) H_{max}.*

2004.

[22] C. E. SHANNON, *A mathematical theory of communication*, The Bell Systems Technical Journal, 27 (1948), pp. 379–423.

[23] J. E. SHORE AND R. W. JOHNSON, *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy*, IEEE Transactions on Information Theory, 26 (1980), pp. 26–36.

[24] D. S. SIVIA, *Data Analysis: A Bayesian Tutorial*, Oxford University Press, Oxford, 1996.

[25] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, N.J., 1973.

[26] N. Sukumar, *Construction of polygonal interpolants: A maximum entropy approach*, International Journal for Numerical Methods in Engineering, 61 (2004), pp. 2159–2181.

[27] ———, *Maximum entropy approximation*, AIP Conference Proceedings, 803 (2005), pp. 337–344.

[28] N. Sukumar and R. W. Wright, *Overview and construction of meshfree basis functions: From moving least squares to entropy approximants*, International Journal for Numerical Methods in Engineering, 70 (2007), pp. 181–205.

[29] D. Walkup and R. J-B Wets, *A Lipschitzian characterization of convex polyhedra*, Proceedings American Mathematical Society, 23 (1969), pp. 167–173.

[30] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, Cambridge, UK, 2005.

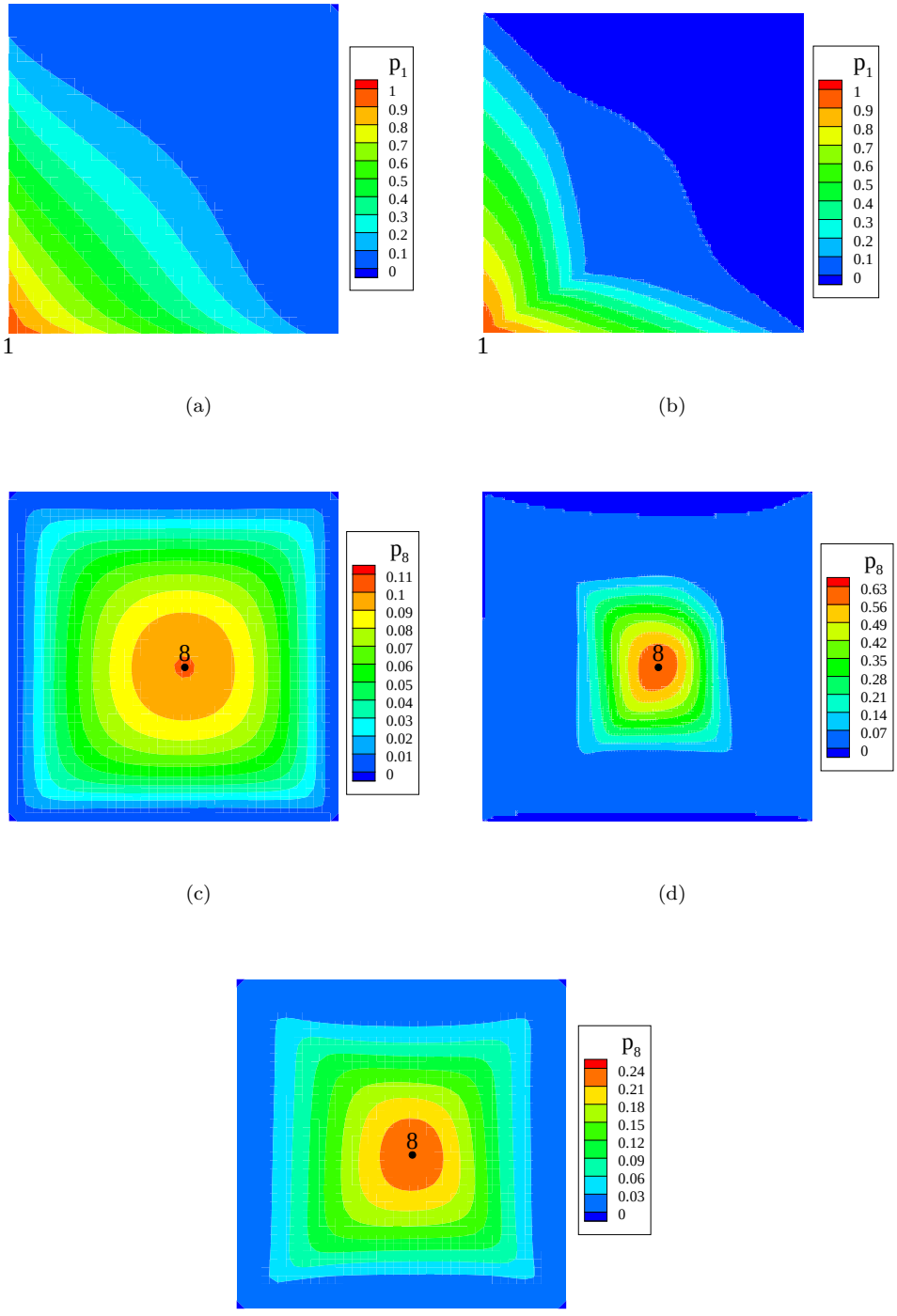


FIG. 3.3. *Two-dimensional entropy basis functions within a unit square. (a),(b)* $p_1(x)$ *with a uniform prior and a Gaussian prior* ($\beta = 20$)*; (c),(d)* $p_8(x)$ *with a uniform prior and a Gaussian prior* ($\beta = 20$)*; and (e)* $p_8(x)$ *with a compactly-supported* C^2 *radial basis function.*